

Heavy-traffic analysis through uniform acceleration of queues with diminishing populations

Gianmarco Bet, Remco van der Hofstad, and Johan S.H. van Leeuwen

December 1, 2015

Abstract

We consider a single server queue that serves a finite population of n customers that will enter the queue (require service) only once, also known as the $\Delta_{(i)}/G/1$ queue. This paper presents a method for analyzing heavy-traffic behavior by using uniform acceleration, which simultaneously lets n and the service rate grow large, while the initial resource utilization approaches one. A key feature of the model is that, as time progresses, more customers have joined the queue, and fewer customers can potentially join. This *diminishing population* gives rise to a class of reflected stochastic processes that vanish over time, and hence do not have a stationary distribution. We establish that, when the arrival times are exponentially distributed, by suitably rescaling space and time, the queue length process converges to a Brownian motion with parabolic drift, a stochastic-process limit that captures the effect of a diminishing population by a negative quadratic drift. When the arrival times are generally distributed, our techniques provide information on the typical queue length and the first busy period.

1 Introduction

The queueing literature is to a large extent built on the assumption that arrivals are governed by some renewal process, a mathematically convenient assumption as it allows the full use of probabilistic tools based on regenerative processes and ergodic theory. This paper, however, considers a *transitory* queueing model, known in the literature as the $\Delta_{(i)}/G/1$ queue, which operates only a finite amount of time and cannot be viewed as a standard regenerative process. The $\Delta_{(i)}/G/1$ queue assumes a finite population of customers entering the queue only once. As time progresses, more customers have joined the queue, and fewer customers can potentially join. This modelling assumption of a *diminishing population of customers* gives rise to a class of reflected stochastic processes that lack a stationary distribution, and instead display relevant behavior only during a *finite time window*. Therefore, only the time-dependent behavior is of interest.

When we assume the arrival times of the customers to be sampled independently from an identical distribution, the arrival times are the order statistics of the sample, and the interarrival times are differences of order statistics. Further assuming a single server, and generally distributed independent service times, this model was coined the $\Delta_{(i)}/G/1$ queue in [17], in which fluid and diffusion limits for the queue length process were established. In [18] a wider class of transitory queues was introduced, with the $\Delta_{(i)}/G/1$ queue still as the prime example, and stochastic-process limits were established for large population sizes.

We will introduce a heavy-traffic regime for the $\Delta_{(i)}/G/1$ queue, leading to stochastic-process limits and heavy-traffic approximations. Considering queueing processes in their critical regimes typically leads to a reduction in complexity, since the complicated processes can often be shown to converge to much simpler limiting stochastic processes. Stochastic-process limits have been studied for single-server queues that have a time-varying arrival rate. Newell [31, 32] pioneered this direction by deriving diffusion approximations, see also [24, 26]. Rigorous results in terms of stochastic-process limits were obtained in [27] (building on [29, 30]). Here, stochastic-process limits were established as refinements to deterministic ODE limits for the time-dependent $M/M/1$ queue, known as the $M_t/M_t/1$ queue. See also [38] for a systematic treatment of the $M_t/G/1$ queue. The technique used in [27] to develop Functional Law of Large Numbers (FLLN) and Functional Central Limit Theorem (FCLT) results uses strong approximation and what is known as the *uniform acceleration* (UA) technique. UA relies on the assumption that the relevant time scale for changes in the queue length process is of the order $O(1/\varepsilon)$ for some $\varepsilon > 0$. Accelerating the process in a uniform manner by scaling the arrival and service rates by ε then reveals the dominant model behavior as $\varepsilon \rightarrow 0$. While in [17, 18] the arrival and service rates are scaled in a similar manner, the time scale considered is of the order $O(1)$. In particular, the time of the process is not scaled. The UA technique has been extensively applied to non-stationary queueing systems with non-homogeneous Poisson input, but it remained unclear whether it is also useful for transitory or time vanishing queueing models as considered in [18] and in the present paper. We will show how the key idea behind the UA technique can be applied to these models. We shall now explain our approach in terms of the easiest setting, in which the identical distribution that generates the arrival times is exponential.

Assume a finite population of n customers, with n very large, and where each customer has an independent exponential clock with mean $1/\lambda$. Customers join the queue when their clocks ring. The initial arrival process (close to time zero) is then roughly Poisson with rate $n\lambda$. However, as time progresses, the arrival intensity decreases, due to those customers that have left the system. Thus, the arrival process is a Poisson process that is thinned according to some time-dependent rule. Denote the mean service time by $1/\mu$. In order to create heavy-traffic conditions we let the population size n grow to infinity, while at the same time making sure that the (initial) traffic intensity $\rho_n = n\lambda/\mu$ is close to one. The system can initially be underloaded (when $n\lambda < \mu$), overloaded (when $n\lambda > \mu$), or critically-loaded (when $n\lambda = \mu$). In case $n\lambda > \mu$, the queue initially shows a roughly linear increase and therefore the correct scaling of the queue length to obtain meaningful limits is n (FLLN) for a first order approximation and $n^{1/2}$ (FCLT) for second order approximations. In particular, no time scaling is needed to obtain these approximations; these are the most relevant approximations obtained in [17].

We focus on the critically-loaded regime, and we combine this with uniform acceleration through the population size n . In the spirit of UA, we let the arrival and service rates scale with n while also rescaling time so as to observe the queue length process at a time scale of order $O(1/n^\gamma)$ for some $\gamma > 0$. Denote the density of the arrival distribution as $f_T(t)$ (with $f_T(t) = \lambda e^{-\lambda t}$, $t \geq 0$ as an important special case), and denote the i.i.d. service requirements of consecutive customers by S_1, S_2, \dots with generic random variable S . Assuming a service rate of n , the service times of consecutive customers are then $D_1 = S_1/n, D_2 = S_2/n, \dots$ with generic random variable D . For the sake of clarity, we now first give a simplified version of the more general heavy-traffic condition (1.2). The heavy-traffic regime we consider is given by the

condition

$$\rho_n := n f_T(0) \mathbb{E}[D] = f_T(0) \mathbb{E}[S] = 1, \quad \text{for } n \text{ large.} \quad (1.1)$$

For the exponential case $f_T(0) = \lambda$, and the condition reads $\rho_n = \lambda \mathbb{E}[S] = 1$ can be interpreted as, for times close to zero, the expected number of newly arriving customer during one service time is roughly one. For general service times, the condition can be understood by interpreting $f_T(t)$ as the *instantaneous arrival rate* in t . Since we consider time scales of the order $O(1/n^\gamma)$, only the mass in zero $f_T(0)$ matters for describing the new arrivals.

In what follows we actually consider a slightly more general definition of the random variables $(D_i)_{i=1}^\infty$ and this leads to the more precise critical scaling

$$\rho_n = 1 + \beta n^{-1/3}. \quad (1.2)$$

The additional term $\beta n^{-1/3}$ arises from detailed calculations, but can be interpreted as the factor that describes the onset of the heavy-traffic period: when $\beta > 0$ (resp. < 0) the queue is initially slightly overloaded (resp. underloaded).

It is clear that the $\Delta_{(i)}/G/1$ queue is strongly influenced by the service-time distribution. In particular, the heavy-traffic behavior is crucially different depending on whether the second moment of the service time distribution is finite or not. In this paper we assume throughout that $\mathbb{E}[S^2] < \infty$, in which case the queueing process is in the domain of attraction of Brownian motion. Indeed, we will take our process and scale space and time so that the stochastic-process limit turns out to be a (reflected) Brownian motion with parabolic drift. The latter process is defined as $(W_t)_{t \geq 0} = (at + bt^2/2 + cB_t)_{t \geq 0}$ with $(B_t)_{t \geq 0}$ a standard Brownian motion, and a, b, c constants. The constant b is negative, so that eventually the free process $(W_t)_{t \geq 0}$ drifts to minus infinity according to $bt^2/2$, causing the reflected process to be essentially stuck at zero. This effect is due to the *diminishing population* effect. One could interpret the quadratic term in the limit as the (cumulative) effect of the customers already served not being able to join the queue again. The stochastic-process limit provides insight into the macroscopic behavior (for n large) of the transitory queueing process, and the different phenomena occurring at different space-time scales. It also gives insight into the orders of the average queue lengths, the probability of large queue lengths occurring, and the time scales of busy periods.

1.1 Comparison with known results

To create the right circumstances for non-degenerate limiting behavior in heavy traffic, we work not only under the critical-load assumption (1.1), but also under the additional assumption that the maximum of the density $f_T(t)$, $t \geq 0$ is assumed in $t = 0$. Indeed, $f_T(t)$ should be interpreted as an *instantaneous arrival rate*, and then $\max_t f_T(t) \mathbb{E}[S] = f_T(t_{\max}) \mathbb{E}[S] = 1$ means that at the peak hour t_{\max} , the queue is critical. While this assumption is trivially satisfied in the representative case of exponential arrivals, it turns out to be of crucial importance when dealing with general arrival times. Let us explain this subtle point in more detail below, and at the same time draw a comparison with existing lines of related research.

In [17, 26], the $\Delta_{(i)}/G/1$ queue is considered, for $t_{\max} > 0$ and a wide range of possible system behaviors related to the maximum $f_T(t_{\max})$. It seems that all the different operating regimes studied in the literature so far can be classified in terms of the parameter $f_T(t_{\max}) \mathbb{E}[S]$ and whether this quantity is smaller (underloaded), equal (heavy traffic), or larger (overloaded) than 1. In fact, even when $f_T(t_{\max}) \mathbb{E}[S] \approx 1$, additional regimes can be considered by changing the rate of convergence (in our case this has to be $1 + \beta/n^{-1/3}$).

In [17] the following result was obtained for the overloaded $\Delta_{(i)}/G/1$ queue:

Theorem (Diffusion limit for overloaded $\Delta_{(i)}/G/1$ queue [17]). *Let $\lim_{n \rightarrow \infty} \bar{Q}_n^\Delta(t)/n = \bar{Q}^\Delta(t)$ be the fluid limit of the queue length process $Q_n^\Delta(t)$. Let*

$$\hat{Q}_n^\Delta(t) = (Q_n^\Delta(t) - n\bar{Q}_n^\Delta(t))/\sqrt{n} \quad (1.3)$$

be the diffusion-scaled queue length process. Then, as $n \rightarrow \infty$,

$$\hat{Q}_n^\Delta(\cdot) \xrightarrow{d} \hat{Q}^\Delta(\cdot), \quad (1.4)$$

with $\hat{Q}^\Delta(\cdot)$ some stochastic process.

The process $\hat{Q}_n^\Delta(\cdot)$ is a diffusion that switches between three regimes: a free Brownian motion, a reflected Brownian motion and the zero process. From the definition of \hat{Q}_n^Δ in [17] it follows easily that, if $f_T(t) \neq f_T(t_{\max})$, the constant $\max_{t \geq 0} \bar{Q}^\Delta(t)$ and the random variable $\max_{t \geq 0} \hat{Q}_n^\Delta(t)$ are different from zero if and only if

$$f_T(t_{\max})\mathbb{E}[S] > 1. \quad (1.5)$$

Therefore, the queue length grows linearly around time t_{\max} (i.e. $\max_{t \geq 0} \bar{Q}^\Delta(t) > 0$) if and only if $f_T(t_{\max})\mathbb{E}[S] > 1$ (overloaded) and the condition $f_T(t_{\max})\mathbb{E}[S] = 1$ is the threshold between the overloaded and underloaded regimes. Hence, $f_T(t_{\max})\mathbb{E}[S] = 1$ defines the critical regime in which the queue grows like a non-trivial power of n , so not linearly.

Louchard [26] considers an approximation of the $\Delta_{(i)}/G/1$ queue, defined as follows. Let $S_n(t)$ be a renewal process associated with the service time random variables $(S_i)_{i=1}^n$. Further, let $A_n(t) = \sum_{i=1}^n \mathbb{1}_{\{T_i \leq t\}}$ be the usual arrival process. Then the queue length process in [26] is defined as

$$Q'_n(t) = \phi(J_n(t)), \quad (1.6)$$

where $J_n(t) := A_n(t) - S_n(t)$ and $\phi(\cdot)$ is the reflection mapping. This is the so-called Borovkov modified system, an approximation in which the server never idles and when a customer arrives and finds an empty queue, his service time is defined to be the remaining time before the next jump time of $S_n(\cdot)$. Louchard [26] then invokes a result by Iglehart and Whitt [19] to argue that this approximation converges to the $\Delta_{(i)}/G/1$ queue length process in *heavy traffic*. For this system, Louchard [26] studies a wide range of system behaviors, by dividing time into intervals that are associated with specific assumptions on $f_T(t)$ and $\mathbb{E}[S]$, and then establishing convergence results within each interval. In that way Louchard [26] identifies using mostly heuristic arguments the possible behaviors of the model. Our setting would then correspond to Cases 2 and 3 in [26] with $\alpha = 1/3$. Hence, while Louchard [26] identifies many possible stochastic-process limits, we exclusively focus on the critical system behavior, and it is that one specific heavy-traffic regime for which we formally derive the stochastic-process limits. Moreover, we also identify further conditions on the density f_T that influence limiting behavior. We show, for instance, that different stochastic-process limits arise depending on how many derivatives of the density $f_T(t)$ in t_{\max} are zero. Although in the first part of the paper we focus on the exponential case, for which all derivatives are different from zero, our methodology can be used to obtain process limits for all such scenarios, as discussed in Section 3.2.

Hence, to force non-degenerate limiting behavior in the critical $\Delta_{(i)}/GI/1$ queue, uniform acceleration is required, in combination with additional assumptions on the arrival times density

in its maximum. This approach also has connections with the work of Mandelbaum and Massey [27] for the $M_t/M_t/1$ queue, who derive a fluid approximation through a FLLN and use this approximation to classify various operating regimes. In this setting, our result corresponds to the ‘Onset of Critical Loading’ regime [27, Theorem 3.4] and the results of [17] correspond to the FLLN and the FCLT [27, Theorems 2.1 and 2.2], as can be seen in Table 1.

model	$\Delta_{(i)}/G/1$	$\Delta_{(i)}/G/1$ [17, 26]	$M_t/M_t/1$ [27]	$M_t/M_t/1$ [27]
regime	crit. loaded in t_{max}	overloaded	crit. loaded in t_{max}	overloaded
time scaling	$n^{-1/3}$	—	$\varepsilon^{1/3}$	—
spatial scaling	$n^{-1/3}$	n^{-1} (FLLN) $n^{-1/2}$ (FCLT)	$\varepsilon^{1/3}$	ε (FLLN) $\varepsilon^{1/2}$ (FCLT)
fluid limit	—	$\phi(\int_0^t (f_T(s) - \mu)ds)$	—	$\phi(\int_0^t (\lambda(s) - \mu(s))ds)$
diffusion limit	$\phi(W)(t)$	regime switching $\begin{cases} B(t) \\ \phi(B)(t) \\ 0 \end{cases}$	$\phi(W)(t)$	regime switching $\begin{cases} B(t) \\ \phi(B)(t) \\ 0 \end{cases}$

Table 1: A comparison between the results for our model, the $\Delta_{(i)}/G/1$ queue and the $M_t/M_t/1$ queue. For simplicity, we have omitted the variances of the Brownian motions.

In Table 1, $W(t) = B(t) + \frac{f'_T(t_{max})}{2}t^2$, $f_T(t)$ denotes the probability density function of the arrival random variables T and $\phi(\cdot)$ is the reflection mapping. Moreover, for the $M_t/M_t/1$ queue it is assumed that the arrival and service processes are inhomogeneous Poisson processes with time-dependent rate $\lambda(s)$ and $\mu(s)$, respectively. The results for the critically loaded $\Delta_{(i)}/G/1$ queue appear in [26] (heuristically) as well as in this paper. It can also be seen from Table 1 that the $M_t/M_t/1$ queue and the $\Delta_{(i)}/G/1$ queue are intimately related. To see this, consider an inhomogeneous Poisson process $N(\int_0^t \lambda(s)ds)$ (with $N(\cdot)$ a Poisson process of rate one). Then, conditioned on $\{N(\int_0^t \lambda(s)ds) = k\}$, the k points themselves are a family of i.i.d. random variables with density given by $\lambda(s)/\int_0^t \lambda(z)dz$. In particular, if $\int_0^\infty \lambda(z)dz < \infty$, conditioned on the total number of Poisson points in $[0, \infty)$ (say, n), the points are a family of n i.i.d. random variables. In [18], the authors explore this connection in relation to the $\Delta_{(i)}/G/1$ queue in great detail.

2 Main results

Let us now summarize our results. We begin by focusing on the case of exponentially distributed arrival times. Given a process $X(\cdot)$ with trajectories in $\mathcal{D}([0, \infty), \mathbb{R})$, the space of càdlàg functions (endowed with the Skorohod J_1 topology), we denote by $\phi(X)(\cdot)$ its reflected version, obtained through the map

$$\phi(f)(x) := f(x) - \inf_{y \leq x} (f \wedge 0)(y). \quad (2.1)$$

The critical behavior of the $\Delta_{(i)}/G/1$ queue is then determined by the following theorem:

Theorem 1 (The critically loaded $\Delta_{(i)}/G/1$ queue with exponential arrivals). *Let $Q_n^\Delta(\cdot)$ be the queue length process of $\Delta_{(i)}/G/1$ queue with rate λ exponential arrivals and service times $(S_i)_{i \geq 1}$ such that $\mathbb{E}[S^2] < \infty$. Assume that the heavy-traffic condition (1.2) holds. Then*

$$n^{-1/3}Q_n^\Delta(\cdot n^{-1/3}) \xrightarrow{d} \phi(W)(\cdot), \quad (2.2)$$

where $W(\cdot)$ is the diffusion process

$$W(t) = \beta\lambda t - \frac{\lambda^2}{2}t^2 + \sigma B(t), \quad (2.3)$$

with $\sigma^2 = \lambda^3 \mathbb{E}[S^2]$ and $B(\cdot)$ a standard Brownian motion.

As an immediate consequence of Theorem 1 we get an asymptotic result for the BP_n , the length of the first busy period in the $\Delta_{(i)}/G/1$ queue (so BP_n describes the number of customers that are served without the queue becoming empty). In order to obtain a sizeable first busy period, we assume that the queue length at time zero is deterministic and grows with n as

$$\lim_{n \rightarrow \infty} \frac{Q_n^\Delta(0)}{n^{1/3}} = q > 0. \quad (2.4)$$

As the next theorem shows, the size of the first busy period depends crucially both on β and q :

Theorem 2 (Busy period of the critical $\Delta_{(i)}/G/1$ queue with exponential arrivals). *Let $Q_n^\Delta(\cdot)$ be the queue length process of the $\Delta_{(i)}/G/1$ queue. Assume that the heavy-traffic condition (1.2) holds. Assume further that (2.4) holds. Let BP_n denote the first busy period of $Q_n^\Delta(\cdot)$. Then*

$$n^{1/3}\text{BP}_n \xrightarrow{d} T_{W_q}^{\beta\lambda}(0), \quad (2.5)$$

where $T_{W_q}^{\beta\lambda}(0)$ is the time until the process $W_q(\cdot)$ crosses level 0, with

$$W_q(t) = q + \beta\lambda t - \frac{\lambda^2}{2}t^2 + \sigma B(t). \quad (2.6)$$

As above, $\sigma^2 = \lambda^3 \mathbb{E}[S^2]$ and $B(\cdot)$ is a standard Brownian motion.

Proof. For a function $f(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ such that $f(0) > 0$, let $T_f(0) := \inf\{t > 0 : f(t) \leq 0\}$ denote the time until $f(\cdot)$ crosses level 0. By Theorem 1 and (2.4),

$$n^{-1/3}Q_n^\Delta(\cdot n^{-1/3}) \xrightarrow{d} W_q(t). \quad (2.7)$$

The functional $f \mapsto T_f(0)$ is a.s. continuous in $W_q(\cdot)$ by [21, Chapter VI, Proposition 2.11]. Moreover,

$$n^{1/3}\text{BP}_n = n^{1/3} \inf\{t > 0 : Q_n^\Delta(t) \leq 0\} = \inf\{t > 0 : n^{-1/3}Q_n^\Delta(tn^{-1/3}) \leq 0\}. \quad (2.8)$$

The conclusion then immediately follows from these observations and the Continuous Mapping Theorem. \square

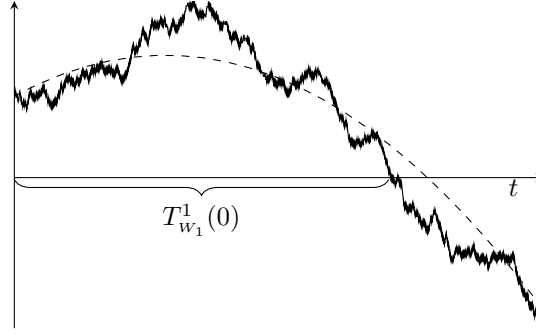


Figure 1: A sample path of $W_1(t)$ with $\beta = 1$ (solid) and the drift $1 + t - \frac{1}{2}t^2$ (dashed).

2.1 Numerical examples

Theorem 2 can be used to obtain numerical approximations of quantities related to the first busy period of the critically loaded $\Delta_{(i)}/G/1$ queue. In [28] an explicit expression for the density of $T_W^\beta(0)$ is given, and numerical simulations are carried out to display how the shape of the density crucially depends on the two parameters q and β . Let $\text{Ai}(x)$ and $\text{Bi}(x)$ denote the classical Airy functions (see [1]).

Theorem 3 (First passage time density [28]). *The first crossing time of zero of $W_q(t) = q + \beta t - 1/2t^2 + \sigma B(t)$ has probability density*

$$f_q(t; \beta, \sigma) = e^{-((t-\beta)^3 + \beta^3)/6\sigma^2 - \beta x} \int_{-\infty}^{+\infty} e^{tu} \frac{B(u)A(u-x) - A(u)B(u-x)}{\pi(A^2(u) + B^2(u))} du, \quad (2.9)$$

where $c = (2\sigma^2)^{1/3}$, $x = q/\sigma^2 > 0$, $A(u) = \text{Ai}(cu)$ and $B(u) = \text{Bi}(cu)$.

Figure 2 shows the convergence of the empirical density function of the first excursion length of a $\Delta_{(i)}/G/1$ queue to the analytic expression provided by [28] and also illustrates the influence of the parameters β and q .

Figure 2 suggests that to obtain a considerable first busy period, both parameters β and q must be chosen appropriately in order to avoid a concentration of the probability mass close to zero. Indeed, in [28] it is conjectured that there exists a \bar{q} such that for all $q > \bar{q}$ and every choice of β the distribution function is unimodal, while for $q < \bar{q}$ there exists a $\bar{\beta} = \bar{\beta}(q)$ such that for $\beta < \bar{\beta}$ the distribution is unimodal, and bimodal otherwise.

In [16] asymptotic expressions are determined for the tail probabilities $\mathbb{P}(T_{W_q}^\beta(0) > x)$ for x large. The most relevant result is the following:

Theorem 4 (Tail busy period length for large x [16]). *For bounded $q/\sigma^2 > 0$,*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{BP}_n \geq xn^{2/3}) = \frac{3\sigma\sqrt{x}}{\sqrt{2\pi}} \frac{\exp(-\mathcal{F}_3(x)/6\sigma^2)}{\mathcal{F}_3'(x)} (1 + O(1/x)), \quad x \rightarrow \infty, \quad (2.10)$$

with $\mathcal{F}_3(x) = (x - \beta)^3 - \frac{1}{4}x^3 - 3qx + \beta^3 + 6\beta q$.

The expression in (2.10) can be used, for example, to study the dependence of the probability of a very large first busy period on the parameters q and β .

We conclude by showing in Table 2 numerical values for the mean busy period for exponential clock times with mean 1 and different values of q and β . Observe that the approximation $\mathbb{E}[\text{BP}_n] \approx n^{-1/3}\mathbb{E}[T_W^\beta(0)]$ is accurate, also for moderate values of n .

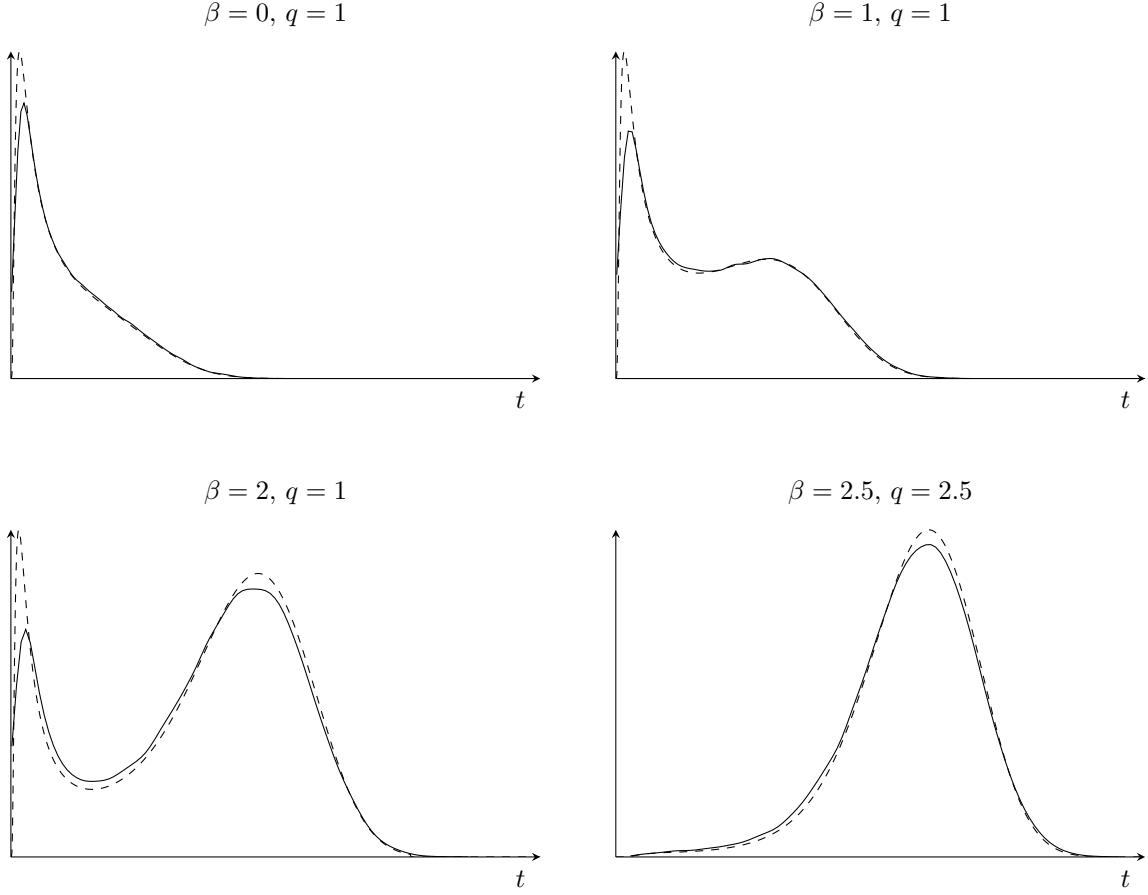


Figure 2: Density plot (dashed line) and Gaussian kernel density estimate (thick line) obtained by running 100000 simulations of a $\Delta_{(i)}/G/1$ queue with $n = 10000$ customers.

n	$q = 1, \beta = 1$		$q = 2, \beta = 1$	
	$n^{1/3}\mathbb{E}[\text{BP}_n]$	rel. error	$n^{1/3}\mathbb{E}[\text{BP}_n]$	rel. error
10	3.0201	0.5072	4.0407	0.4079
100	2.2170	0.1062	3.2611	0.1362
1000	2.0341	0.0151	2.9813	0.0387
10000	2.0306	0.0133	2.9351	0.0226
100000	2.0295	0.0128	2.9145	0.0155
∞	2.0038	—	2.8701	—

Table 2: Mean busy period for the pre-limit queue with different population sizes and the exact expression for $n = \infty$ computed using (2.9). Each value for the pre-limit queue is the average of 10000 simulations.

2.2 Outline

The remainder of the paper is devoted to proving Theorems 1 and 2. The proof of Theorem 1 consists of several steps. In Section 3 we give a detailed overview of the proof. In Section 4.1 we settle some preliminaries and the required notation. Briefly, the proof then proceeds as

follows. First we approximate the $\Delta_{(i)}/G/1$ queue via a certain process $Q_n(\cdot)$ and this is shown to converge when appropriately rescaled (Section 4). The approximation is then shown to be asymptotically equivalent to the $\Delta_{(i)}/G/1$ queue (Section 7). As it turns out, the process $Q_n(\cdot)$ converges under much more general assumptions on the density in zero of the arrival clocks (Section 5 and Section 6). While it does not seem to be possible to use this result to prove process convergence of the $\Delta_{(i)}/G/1$ queue process for more general arrivals, it can be exploited to provide insight into the behavior of its first busy period, and on the typical queue length (Section 3.2). In Section 8 we discuss various connections with the literature in queueing theory, random graphs and statistics. There, we also discuss several future research directions.

3 Overview of the proof

The proof of Theorem 1 proceeds in several steps. First, we construct a queueing model that *approximates* the $\Delta_{(i)}/G/1$ queue. Second, we show that it *coincides* with the $\Delta_{(i)}/G/1$ queue when embedded after service completions. Here we crucially use the properties of the exponential distribution. Third, we prove a version of Theorem 1 for the (embedded) approximating model. Lastly, we show via a time change argument that the $\Delta_{(i)}/G/1$ queue and its embedded counterpart converge to the same limit.

Let us now present some more details about the approximating model we consider. Like in the $\Delta_{(i)}/G/1$ queue, we consider n customers in a population that all possess independent clocks that ring after i.i.d. clock times with exponential random variable T . Whenever a clock rings, that customer joins the queue. Customers are served in order of arrival. The service requirements of consecutive customers are given by the i.i.d. random variables $(S_i)_{i=1}^n$, independent from n . We assume that $\mathbb{E}[S^2] < \infty$. We further assume that the service capacity per time unit scales as $\frac{1+\beta n^{-1/3}}{n}$, so that the service times are given by

$$D_i = \frac{S_i}{n} \left(1 + \beta n^{-1/3}\right), \quad i = 1, \dots, n. \quad (3.1)$$

After service completion, the customer is removed from the system. As explained below (1.1) we shall work under the heavy-traffic condition

$$\rho_n = n\lambda\mathbb{E}[D] = 1 + \beta n^{-1/3}. \quad (3.2)$$

If, after a service completion, the system is empty, the customer with the smallest arrival time is drawn from the population and is immediately put into service. As will become clear, considering the queue length process embedded at service completions makes the process more amenable to mathematical analysis (e.g. allowing access to discrete-time martingale techniques). Let $Q_n(k)$ denote the number of customers in the queue just after the service completion of the k -th customer. The queue length process $(Q_n(k))_{k \geq 0}$, embedded at service completions, is then given by $Q_n(0) = 0$ and

$$Q_n(k) = (Q_n(k-1) + A_n(k) - 1)^+, \quad k = 1, 2, \dots \quad (3.3)$$

with $x^+ = \max\{0, x\}$ and $A_n(k)$ the number of arrivals during the service time of the k -th customer. Here $A_n(k)$ is given by

$$A_n(k) = \sum_{i \notin \nu_k} \mathbf{1}_{\{\sum_{j=1}^{k-1} D_j \leq T_i \leq \sum_{j=1}^k D_j\}} \quad (3.4)$$

where ν_k is the set of customers no longer in the population at the beginning of the service of the k -th customer. The system defined in (3.3) and (3.4) neglects idle times, which is a simplification of the $\Delta_{(i)}/G/1$ model. This in turn will greatly simplify the analysis since it allows for the representation of the process as (3.3) and (3.4).

It is possible to give an equivalent definition of the process $Q_n(k)$ in (3.3) through the reflection map. Define the process $(N_n(k))_{k \geq 0}$ by $N_n(0) = 0$ and

$$N_n(k) = N_n(k-1) + A_n(k) - 1. \quad (3.5)$$

Then it is easy to see that

$$(Q_n(k))_{k \geq 0} = (\phi(N_n)(k))_{k \geq 0} \quad \text{w.p. 1.}$$

To avoid unnecessary notation, for all discrete-time processes $(X(k))_{k \geq 0}$, we write $X(t)$, with $t \in \mathbb{R}^+$ instead of $X(\lfloor t \rfloor)$. Note that the process defined in (3.5) may take negative values. We first prove a limit theorem for $N_n(\cdot)$ and then apply the reflection map to obtain a limit for the queue length process $Q_n(\cdot)$ defined in (3.3). The critical behavior of this model is determined in the following theorem:

Theorem 5 (Convergence of the approximating process). *Let $Q_n(\cdot)$ be the process defined in (3.3) and (3.4), associated with rate λ exponential arrival random variables $(T_i)_{i \geq 1}$ and service times $(S_i)_{i \geq 1}$ such that $\mathbb{E}[S^2] < \infty$. Then*

$$n^{-1/3} N_n(\cdot n^{2/3}) \xrightarrow{d} W(\cdot), \quad (3.6)$$

where $W(\cdot)$ is the diffusion process

$$W(t) = \beta t - \frac{1}{2} t^2 + \sigma B(t), \quad (3.7)$$

with $\sigma^2 = \lambda^2 \mathbb{E}[S^2]$ and $B(\cdot)$ a standard Brownian motion. Moreover,

$$n^{-1/3} Q_n(\cdot n^{2/3}) \xrightarrow{d} \phi(W)(\cdot). \quad (3.8)$$

Recall that $\phi(W)(\cdot)$ denotes the reflected version of $W(\cdot)$. The time-scaling exponent $2/3$ that appears in Theorem 6 is intimately related to the exponent that appears in Table 1. By (3.4), the clocks T_i of the customers joining the queue are of the order $T_i = O_{\mathbb{P}}(1/n)$. Thus, after scaling time by $n^{2/3}$ we effectively observe the queue at times of the order $O(n^{-1/3})$.

We now provide a heuristic argument that explains the scaling exponents in Theorem 6. With $\Sigma_i := \sum_{l=1}^i D_l/n$,

$$\begin{aligned} Q_n(tn^\alpha) &\approx \sum_{i=1}^{tn^\alpha} \left(\sum_{j \notin \nu_i} \mathbb{1}_{\{\sum_{l=1}^{i-1} D_l \leq T_j \leq \sum_{l=1}^i D_l\}} - 1 \right) \\ &\approx \sum_{i=1}^{tn^\alpha} (n(F_T(\Sigma_i) - F_T(\Sigma_{i-1})) - 1) \\ &\approx \sum_{i=1}^{tn^\alpha} (f_T(\Sigma_{i-1})\mathbb{E}[S] - 1) \approx \sum_{i=1}^{tn^\alpha} \sum_{l=1}^{i-1} \frac{S_l}{n} f'_T(0)\mathbb{E}[S], \end{aligned} \quad (3.9)$$

where the last approximation comes from (1.1). This computation gives us the leading order term of the queue length process up to a multiplicative constant

$$Q_n(tn^\alpha) \approx \sum_{i=1}^{tn^\alpha} \frac{i}{n} \approx \frac{t^2}{2} n^{2\alpha-1}. \quad (3.10)$$

The queue is the sum of (the order of) n^α contributions, thus (ignoring dependencies) the correct spatial scaling in order to obtain Gaussian fluctuations is $n^{\alpha/2}$. Then, in order to obtain both a deterministic drift and a Brownian contribution in the limit, the order of magnitude of the first order approximation (3.9) should equate that of the diffusion approximation, $n^{\alpha/2}$. Hence $2\alpha - 1 = \alpha/2$, so that α should be $2/3$.

3.1 Overview of the proof of Theorem 1

We now show how Theorem 1 can be deduced from Theorem 5 through a time change argument. With $k \mapsto \tilde{Q}_n^\Delta(k)$ denoting the $\Delta_{(i)}/G/1$ queue embedded at service completions, we first argue that $\tilde{Q}_n^\Delta(\cdot)$ is closely related to Q_n :

Lemma 1 (Distribution of the embedded $\Delta_{(i)}/G/1$ queue). *For all $k \geq 1$,*

$$Q_n(k) \stackrel{d}{=} \tilde{Q}_n^\Delta(k). \quad (3.11)$$

Proof. One can see this by coupling the two queues as follows. The sequence of service times $(S_i)_{i=1}^n$ are taken to be the same for the two queues while the arrival clocks coincide until the end of the first busy period. After that, assign new clocks to the customers still in the population. The first customer after the idle period in the $\Delta_{(i)}/G/1$ queue is also the customer placed into service in the approximating model. At the beginning of the busy period, assign new clocks to the customers still in the population. The coupling then proceeds in this manner until the population in both queues is depleted. By the properties of exponentials, these new processes (with the clocks drawn multiple times) coincide in distribution with the original ones (with the clocks drawn at the start of the system), since

$$\mathbb{P}(T_i \geq B + I + x | T_i \geq B + I) = \mathbb{P}(T_i \geq x) = \mathbb{P}(T_i \geq B + x | T_i \geq B), \quad (3.12)$$

where $B = B(t)$ is the busy time process and $I = I(t)$ the idle time process at the instant t in which a new busy period starts. The number of arrivals during one service time is then the same in the two coupled queues because the arrival times are equal. In particular, the queues sampled at the end of a service time have the same distribution. \square

The next step is to prove that the supremum distance between $Q_n(\cdot)$ and $Q_n^\Delta(\cdot)$ (when suitably rescaled in space and time) converges to zero. Let $\|f\|_T := \sup_{t \leq T} |f(t)|$ denote the supremum norm. The claim is contained in the following lemma:

Lemma 2 (Asymptotic equivalence of the approximating model). *For each $T > 0$, as $n \rightarrow \infty$,*

$$n^{-1/3} \left\| Q_n\left(\frac{\cdot}{\mathbb{E}[S]} n^{2/3}\right) - Q_n^\Delta(\cdot n^{-1/3}) \right\|_T \xrightarrow{\mathbb{P}} 0. \quad (3.13)$$

In particular, by Slutsky's theorem, Lemma 2 and Theorem 5 imply

$$n^{-1/3}Q_n^\Delta(\cdot n^{-1/3}) \xrightarrow{d} W(\lambda \cdot), \quad (3.14)$$

where $W(t) = \beta t - 1/2t^2 + B(t)$ and $B(\cdot)$ Brownian motion. Without loss of generality, we will assume from now on that $\mathbb{E}[S] = 1/\lambda = 1$. To prove (3.13) we split

$$\|Q_n(\cdot) - Q_n^\Delta(\cdot)\|_T \leq \|Q_n(\cdot) - Q_n(\varphi_n(\cdot))\|_T + \|Q_n(\varphi_n(\cdot)) - Q_n^\Delta(\cdot)\|_T, \quad (3.15)$$

for an appropriate yet still unspecified time change $\varphi_n(\cdot)$.

Thus, we are left to prove the two following claims:

- (i) $n^{-1/3}\|Q_n(\cdot n^{2/3}) - Q_n(\varphi_n(\cdot)n^{2/3})\|_T \xrightarrow{\mathbb{P}} 0$;
- (ii) $n^{-1/3}\|Q_n(\varphi_n(\cdot)n^{2/3}) - Q_n^\Delta(\cdot n^{-1/3})\|_T \xrightarrow{\mathbb{P}} 0$.

The idea behind introducing $\varphi_n(\cdot)$ is to rescale time so that each time-step (corresponding to one service) is replaced by the actual length of the service time. In this way, the interval $[0, 1]$ is replaced by $[0, S_1/n]$, interval $[1, 2]$ is replaced by $[S_1/n, S_1/n + S_2/n]$ (if a customer has arrived during the first service), and so on. The time change $\varphi_n(\cdot)$ must also take idle times into account. The precise expression of $\varphi_n(\cdot)$ is given in Section 7. Claim (i) then states that the time change $\varphi_n(\cdot)$ is asymptotically trivial. Indeed, the following holds:

Lemma 3 (Verification of Claim (i)). *As $n \rightarrow \infty$,*

$$\|t - \varphi_n(t)\|_T \xrightarrow{\mathbb{P}} 0. \quad (3.16)$$

Consequently,

$$n^{-1/3}\|Q_n(\cdot n^{2/3}) - Q_n(\varphi_n(\cdot)n^{2/3})\|_T \xrightarrow{\mathbb{P}} 0. \quad (3.17)$$

A crucial step in proving (3.16) is in proving that the idle time process of the $\Delta_{(i)}/G/1$ queue is negligible in the scaling regime we consider. We postpone the proof of this and of Lemma 3 to Section 7.

After this time change the two queues $Q_n(\varphi_n(\cdot))$ and $Q_n^\Delta(\cdot)$ are synchronized in time. It still remains to be proven that their supremum distance converges to zero. However, in their coupling $Q_n(\varphi_n(\cdot))$ is constructed by sampling $Q_n^\Delta(\cdot)$ at service completions, so that the two coincide at the time of each service completion. In other words, the maximum distance between $Q_n(\varphi_n(\cdot))$ and $Q_n^\Delta(\cdot)$ is the maximum number of arrivals during a single service time up until time T , that is

$$\|Q_n(\varphi_n(\cdot)n^{2/3}) - Q_n^\Delta(\cdot n^{-1/3})\|_T = \max_{k \leq Tn^{2/3}} A_n(k), \quad (3.18)$$

where $A_n(k)$ is the number of arrivals during the k -th service time, as defined in (3.4). This is proven to be negligible, thus concluding the proof of Lemma 2, in the following lemma:

Lemma 4 (Verification of Claim (ii)). *As $n \rightarrow \infty$,*

$$n^{-1/3} \max_{k \leq Tn^{2/3}} A_n(k) \xrightarrow{\mathbb{P}} 0. \quad (3.19)$$

Consequently,

$$n^{-1/3}\|Q_n(\varphi_n(\cdot)n^{2/3}) - Q_n^\Delta(\cdot n^{-1/3})\|_T \xrightarrow{\mathbb{P}} 0. \quad (3.20)$$

We postpone the proof to Section 7.

3.2 General arrivals

Theorem 5 can be generalized to allow for generally distributed arrival times. Let us now introduce our assumptions for this case.

The arrival times of the n customers are drawn independently from a common distribution. Let the random variable T have distribution function F_T . Assume that T admits a positive and continuous density function $f_T(\cdot)$, with $f_T(0) \in (0, \infty)$. In addition, as a technical assumption, we assume that the sublinear terms of the distribution function F_T decay as quickly as

$$F_T(x) - F_T(\bar{x}) = f_T(\bar{x})(x - \bar{x}) + o(|\bar{x} - x|^{4/3}) \quad \forall \bar{x} \in (0, \infty). \quad (3.21)$$

This is, for example, the case when $F_T \in \mathcal{C}^2([0, \infty))$. As an additional technical assumption, this error should be uniform over \bar{x} , for small values of \bar{x} , as in

$$\sup_{\bar{x} \leq Cy^{1/3}} |F_T(\bar{x} + y) - F_T(\bar{x}) - f_T(\bar{x})y| = o(y), \quad (3.22)$$

where we set $x - \bar{x} = y$ for convenience. We also assume that $f'_T(\cdot)$ exists and is continuous in a neighborhood of zero. In particular (3.22) is always satisfied since

$$\sup_{\bar{x} \leq Cy^{1/3}} |F_T(\bar{x} + y) - F_T(\bar{x}) - f_T(\bar{x})y| = \sup_{\substack{\bar{x} \leq Cy^{1/3} \\ \zeta \in (\bar{x}, \bar{x} + y)}} \left| \frac{f'_T(\zeta)}{2} y^2 \right| \leq \frac{M}{2} y^2, \quad (3.23)$$

where M is the supremum of $f'_T(\cdot)$ in a neighborhood of zero. In particular, it also holds that

$$f_T(x) = f_T(0) + f'_T(0) \cdot x + o(x), \quad (3.24)$$

Since $\lim_{x \rightarrow \infty} f_T(x) = 0$ and f_T is continuous on $[0, \infty)$, it admits a maximum in $[0, \infty)$. While the maximum can be attained in principle in multiple points, a crucial additional assumption is that it is also attained in 0, that is

$$f_T(0) = \sup_{x \geq 0} f_T(x). \quad (3.25)$$

The service requirements of consecutive customers are again given by (3.1) and are such that $\mathbb{E}[S^2] < \infty$. The heavy-traffic regime for this model is given by a condition analogous to (3.2), that is

$$\rho_n = n f_T(0) \cdot \mathbb{E}[D] = 1 + \beta n^{-1/3}. \quad (3.26)$$

Theorem 6 (Convergence of the approximating process for general arrivals). *Let $Q_n(\cdot)$ be the process defined in (3.3) and (3.4), associated with the arrival random variables $(T_i)_{i \geq 1}$ and service times $(S_i)_{i \geq 1}$ such that $\mathbb{E}[S^2] < \infty$. Then, under the assumptions as above,*

$$n^{-1/3} N_n(\cdot n^{2/3}) \xrightarrow{d} W(\cdot), \quad (3.27)$$

where $W(\cdot)$ is the diffusion process

$$W(t) = \beta t + \frac{f'_T(0)}{2f_T(0)^2} t^2 + \sigma B(t), \quad (3.28)$$

with $\sigma^2 = f_T^2(0)\mathbb{E}[S^2]$ and $B(\cdot)$ a standard Brownian motion. Moreover,

$$n^{-1/3} Q_n(\cdot n^{2/3}) \xrightarrow{d} \phi(W)(\cdot). \quad (3.29)$$

We postpone the involved proof to Section 5. Note that when T is exponentially distributed, $f'_T(0)/(2f_T(0)^2) = -1/2$, so that Theorem 6 can be seen as a generalization of Theorem 5. The lack of the memoryless property for the arrival clocks T_i makes that the arguments in Section 3.1 cannot be carried over to this case. Therefore, it does not seem possible to deduce Theorem 1 (for general arrival clocks) from Theorem 6. For example, the coupling between the approximating model and the $\Delta_{(i)}/G/1$ queue in the proof of Lemma 1 will break down after the end of the first busy period. However, this still allows us to prove results for the first busy period of the $\Delta_{(i)}/G/1$ queue with general arrivals. In particular we have the following generalization of Theorem 2:

Theorem 7 (Busy period of the critical $\Delta_{(i)}/G/1$ queue with general arrivals). *Let $Q_n^\Delta(\cdot)$ be the queue length process of the $\Delta_{(i)}/G/1$ queue. Assume that (2.4) holds. Assume further (3.21)-(3.26) and that $f'_T(0) < 0$. Let BP_n denote the first busy period of $Q_n^\Delta(\cdot)$. Then*

$$n^{1/3}\text{BP}_n \xrightarrow{d} T_{W_q}^\beta(0), \quad (3.30)$$

where $T_{W_q, f_T}^\beta(0)$ is the time until the process $W_{q, f_T}(\cdot)$ crosses level 0, with

$$W_{q, f_T}(t) = q + \beta f_T(0)t + \frac{f'_T(0)}{2}t^2 + \sigma B(t). \quad (3.31)$$

Here, $\sigma^2 = f_T^3(0)\mathbb{E}[S^2]$ and $B(\cdot)$ is a standard Brownian motion.

Proof. By coupling the approximating model and the $\Delta_{(i)}/G/1$ queue as in the proof of Lemma 1, we get that

$$Q_n(k) \stackrel{d}{=} \tilde{Q}_n^\Delta(k), \quad \text{for } k \leq T_{Q_n}(0). \quad (3.32)$$

that is, the two embedded processes coincide in distribution until the end of the first busy period. By proceeding as in Lemma 2 we see that

$$n^{-1/3}\|Q_n(\cdot f_T(0)n^{2/3}) - Q_n^\Delta(\cdot n^{-1/3})\|_{T_{Q_n}(0)} \xrightarrow{\mathbb{P}} 0. \quad (3.33)$$

In particular,

$$n^{-1/3}Q_n^\Delta(\cdot \wedge T_{Q_n^\Delta}(0)n^{-1/3}) \xrightarrow{d} W_q((\cdot \wedge T_{W_q}(0))f_T(0)). \quad (3.34)$$

Then,

$$\begin{aligned} n^{1/3}\text{BP}_n &= n^{1/3} \inf\{t > 0 : Q_n^\Delta(t) \leq 0\} \\ &= \inf\{t > 0 : n^{-1/3}Q_n^\Delta(tn^{-1/3}) \leq 0\}. \end{aligned} \quad (3.35)$$

Finally, the same continuity argument as in Theorem 2 gives the conclusion. \square

In Table 3 we show numerically that the (rescaled) average busy period of the $\Delta_{(i)}/G/1$ queue with general (hyperexponential) arrivals converges to the exact value obtained by Lemma 3. The arrival random variable is exponentially distributed with rate $\lambda_1 = 2$ with probability $p_1 = 0.2$ and with rate $\lambda_2 = 3/4$ with probability $p_2 = 0.8$. Note that formula (2.9) holds for a parabolic drift of the form $-\frac{1}{2}t^2$. However, this can be extended to more general coefficients of the parabolic term by some simple scaling properties. In particular, the first hitting time of zero of $W_{q, f_T}(\cdot)$ is distributed as

$$T_{W_q, f_T}^\beta(0) \stackrel{d}{=} k^{-2/3}T_{W_{qk^{1/3}}}^{\beta f_T(0)k^{-1/3}}(0), \quad (3.36)$$

n	$q = 1, \beta = 1$		$q = 2, \beta = 1$	
	$n^{1/3}\mathbb{E}[\text{BP}_n]$	rel. error	$n^{1/3}\mathbb{E}[\text{BP}_n]$	rel. error
10	2.8630	0.6581	3.8646	0.5620
100	1.9862	0.1503	2.9665	0.1991
1000	1.8103	0.0484	2.6486	0.0706
10000	1.7725	0.0265	2.5596	0.0346
100000	1.7440	0.0100	2.5050	0.0125
∞	1.7267	—	2.4740	—

Table 3: Mean busy period for the pre-limit queue with general (hyperexponential) arrivals and different population sizes and the exact expression for $n = \infty$ computed using (2.9). The hyperexponential distribution has parameters $(p_1, p_2, \lambda_1, \lambda_2) = (0.2, 0.8, 2, 3/4)$. Each value for the pre-limit queue is the average of 10000 simulations.

where $k = f'_T(0)$ and $T_{W_q}^\beta(0)$ is defined as below (2.5). Relation (3.36) follows from a more general scaling relation, see e.g. [16, Section 4.1].

Several difficulties arise when trying to prove Theorem 1 for general arrivals using similar techniques as the ones for exponential arrivals. These difficulties can be traced back to the fact that the $\Delta_{(i)}/G/1$ queue and its approximating process cannot be coupled beyond the first busy period, due to the lack of memoryless property of the arrival clocks. However, there are several strong indications that Theorem 1 should also hold for generally distributed arrival clocks. The proof of Theorem 7 implies that the first excursion of the $\Delta_{(i)}/G/1$ queue length process to the first excursion of an appropriate Brownian motion with parabolic drift. Moreover, the limiting process (3.31) only depends on the distribution of T through $f_T(0)$ (λ for exponential clocks), suggesting that the result is insensitive to the arrival clocks distribution, as long as $f_T(0) > 0$. This can be further justified as follows. Let $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$ the order statistics of the arrival clocks $(T_i)_{i=1}^n$. These describe the ordered arrival times of the customers. By a standard argument, $T_{(i)}$ is distributed as

$$T_{(i)} \stackrel{d}{=} F_T^{-1}(1 - \exp(-E_{(i)})), \quad (3.37)$$

where $E_{(i)}$ is the i -th order statistic associated with n i.i.d. exponential random variables E_1, E_2, \dots, E_n with unitary mean. By Taylor expanding expression (3.37) one sees that the order statistics of $(T_i)_{i=1}^n$ are closely approximated by those of $(E_i)_{i=1}^n$. Therefore, the result for exponential arrivals should be closely related to the one for general arrivals. These considerations lead us to formulate the following conjecture:

Conjecture: Theorem 1 holds for generally distributed arrival times satisfying assumptions (3.21)-(3.26), with $W(t) = \beta f_T(0)t - \frac{f_T^{(0)}}{2}t^2 + \sigma B(t)$, where $\sigma^2 = f_T(0)^3 \mathbb{E}[S^2]$ and $B(\cdot)$ a standard Brownian motion.

3.2.1 The ℓ -th order contact case

The technique developed to prove Theorem 6 can be exploited to prove limit results for the more general case in which the function $t \mapsto f_T(t) - 1/\mathbb{E}[S]$ has ℓ -th order contact in zero, defined as follows:

Definition 1 (ℓ -th order contact point). *Given a smooth, real-valued, function $f(t)$, we say it has ℓ -th order contact in \bar{t} if $f(\bar{t}) = 0$, $f^{(l)}(\bar{t}) = 0$ for $l = 1, \dots, \ell - 1$ and $f^{(\ell)}(\bar{t}) \neq 0$.*

If $f(\bar{t}) = o(1)$ and all the other assumptions on Definition 1 are kept, we still say that $f(\cdot)$ has an ℓ -th order contact in \bar{t} . Indeed, our criticality assumption is $f_T(0) - 1/\mathbb{E}[S] = o(1)$, where the error term is specified later. The assumption that the argmax of $f_T(\cdot)$ is zero allows us to consider both odd and even order contacts. In this case, Theorem 6 can be generalized as follows

Theorem 8 (Asymptotics for the critical ℓ -th order contact queue). *Assume that the function $f_T(t) - 1/\mathbb{E}[S]$ has ℓ -th order contact in 0, where $\ell \geq 1$. Define*

$$\alpha = \frac{\ell}{\ell + 1/2}. \quad (3.38)$$

Then,

$$n^{-\alpha/2} Q_n(\cdot n^\alpha) \xrightarrow{d} \phi(W)(\cdot), \quad (3.39)$$

where $W(\cdot)$ is given by

$$W(t) = \beta t - ct^{\ell+1} + \sigma B(t), \quad c, \sigma \in \mathbb{R}^+, \quad (3.40)$$

and $B(\cdot)$ is a standard Brownian motion.

Note that $\ell = 1$, corresponding to the case considered in Theorem 6, returns the same scaling as in Theorem 6. Moreover, the case $\ell = 2$ has already been known in the literature for quite some time, at least at a heuristic level. Newell derived the correct exponents ($\alpha = 4/5$) through an argument using the Fokker-Planck equation associated with the queue length process (see [31, part III]). Notice also that $\lim_{\ell \rightarrow \infty} \frac{\alpha}{2} = \lim_{\ell \rightarrow \infty} \frac{\ell}{2\ell+1} = \frac{1}{2}$, suggesting that the right scaling for the uniform arrivals case (∞ -order contact) is the diffusive one.

An analogous heuristic to the one below 5 motivates the expressions for α and the limit process. We first consider a first-order approximation of the queue. Let $\Sigma_i := \sum_{l=1}^i D_l/n$, then

$$\begin{aligned} Q_n(tn^\alpha) &\approx \sum_{i=1}^{tn^\alpha} \left(\sum_{j \notin \nu_i} \mathbb{1}_{\{\sum_{l=1}^{i-1} D_l \leq T_j \leq \sum_{l=1}^i D_l\}} - 1 \right) \\ &\approx \sum_{i=1}^{tn^\alpha} (n(F_T(\Sigma_i) - F_T(\Sigma_{i-1})) - 1) \\ &\approx \sum_{i=1}^{tn^\alpha} (f_T(\Sigma_{i-1}) \cdot \mathbb{E}[S] - 1) \approx \sum_{i=1}^{tn^\alpha} \left(\sum_{l=1}^{i-1} S_l/n \right)^\ell f_T^{(\ell)}(0) \mathbb{E}[S], \end{aligned} \quad (3.41)$$

where as before the last approximation comes from the criticality assumption. Thus the leading order term (up to a multiplicative constant) of the queue length process is

$$Q_n(tn^\alpha) \approx \sum_{i=1}^{tn^\alpha} \frac{i^\ell}{n^\ell} \approx t^{\ell+1} n^{(\ell+1)\alpha - \ell}. \quad (3.42)$$

For both a deterministic drift and a Brownian contribution to appear in the limit, the order of magnitude of the first order approximation (3.42) should equate $n^{\alpha/2}$. This gives

$$(\ell + 1)\alpha - \ell = \alpha/2 \Rightarrow \alpha = \frac{\ell}{\ell + 1/2}. \quad (3.43)$$

The formal proof of Theorem 8 mimics what has been done for the case $\ell = 1$. One might expect that assumptions on the existence of higher moments of S are needed, but this is not the case. In Section 6, we perform the key steps in the analysis in order to show how to proceed in this general case.

When the arrival clocks distribution function has an ℓ -th order contact in 0, a result analogue to Theorem 7 holds for the $\Delta_{(i)}/G/1$ queue. However, the parabolic drift is replaced by a higher order polynomial drift, and no closed form formulas are available for computing statistics of the first busy period. We leave this issue for future research. In Table 4 we compute the mean first busy period of a $\Delta_{(i)}/G/1$ queue with half-normally distributed arrival clocks. Recall that, if Z is normally distributed with mean 0 and variance σ^2 , then $|Z|$ is half-normally distributed with scale parameter σ . In particular, $f'_T(0) = 0$ and $f''_T(0) < 0$.

n	$q = 1, \beta = 1$	$q = 2, \beta = 1$
	$n^{1/5}\mathbb{E}[\text{BP}_n]$	$n^{1/5}\mathbb{E}[\text{BP}_n]$
10	3.8340	5.3984
100	3.0997	4.1232
1000	2.8378	3.8772
10000	2.7801	3.7721
100000	2.7942	3.7548

Table 4: Mean busy period for the pre-limit queue with general (half-normal) arrivals and different population sizes. The half-normal distribution has scale parameter $\sigma = \sqrt{\pi}/\sqrt{2}$, and $f'_T(0) = 0$, $f''_T(0) < 0$. Each value for the pre-limit queue is the average of 10000 simulations.

3.3 Preliminaries

Here we present some auxiliary results used in the proof of Theorem 6. All random variables defined from now on are defined on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. When needed, elements of Ω will always be denoted by ω . Given two real-valued random variables X, Y we say that X *stochastically dominates* Y , and we denote it by $Y \preceq X$, if

$$\mathbb{P}(X \leq x) \leq \mathbb{P}(Y \leq x), \quad \forall x \in \mathbb{R},$$

so that for every nondecreasing function $f : \mathbb{R} \rightarrow \mathbb{R}$

$$\mathbb{E}[f(Y)] \leq \mathbb{E}[f(X)]. \quad (3.44)$$

If X and Y are defined on the same probability space Ω , and $X(\omega) \leq Y(\omega)$ for almost every $\omega \in \Omega$, then we write $X \stackrel{\text{a.s.}}{\leq} Y$. We write $f(n) = O(g(n))$ for functions $f, g \geq 0$ and $n \rightarrow \infty$ if there exists a constant $C > 0$ such that $\lim_{n \rightarrow \infty} f(n)/g(n) \leq C$, and $f(n) = o(g(n))$ if $\lim_{n \rightarrow \infty} f(n)/g(n) = 0$. Furthermore, we write $O_{\mathbb{P}}(a_n)$ for a sequence of random variables X_n for which $|X_n|/a_n$ is tight as $n \rightarrow \infty$. Moreover, we write $o_{\mathbb{P}}(a_n)$ for a sequence of random

variables X_n for which $|X_n|/a_n \xrightarrow{\mathbb{P}} 0$ as $n \rightarrow \infty$. We say that a sequence of events $(E_n)_{n \geq 1}$ holds with high probability, or w.h.p. for short, if $\mathbb{P}(E_n) \rightarrow 1$ as $n \rightarrow \infty$.

All the processes we deal with are elements of the space $\mathcal{D}([0, \infty), \mathbb{R})$ (or \mathcal{D} , for short) of càdlàg functions, defined as the space of all functions $f : [0, \infty) \rightarrow \mathbb{R}$ such that for all $\bar{x} \in [0, \infty)$

- $\lim_{x \rightarrow \bar{x}^+} f(x) = f(\bar{x})$;
- $\lim_{x \rightarrow \bar{x}^-} f(x) < \infty$.

Hence, $N_n(\cdot)$ is a \mathcal{D} -valued random variable. Following [5], we say that X_n *converges in distribution* (or *weakly converges*) to X (and denote it by $X_n \xrightarrow{d} X$) if $\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]$ as n tends to infinity for every $f(\cdot)$ that is real-valued, bounded and continuous. In particular, if X is \mathcal{D} -valued, $f(\cdot)$ can be any (almost everywhere) continuous function from \mathcal{D} to \mathbb{R} . Thus, to formally establish convergence in distribution in the space of \mathcal{D} -valued random variables, a metric, or at least a topology, on \mathcal{D} is needed (in order to define continuity of the functions $f(\cdot)$). Several topologies on the space \mathcal{D} have been defined (all by Skorokhod, in his famous paper [34]). For our purposes we will consider the J_1 topology, which can be described as being generated by some metric d_∞ on $\mathcal{D}([0, \infty), \mathbb{R})$ defined as an extension of some metric d_t on $\mathcal{D}([0, t], \mathbb{R})$. The latter is defined as follows. Let $\|\cdot\|$ indicate the supremum norm, $\text{id}(\cdot)$ the identity function on $[0, t]$ and Λ_t the space of nondecreasing homeomorphisms on $[0, t]$. Define, for any $x_1, x_2 \in \mathcal{D}$,

$$d_t(x_1, x_2) := \inf_{\lambda \in \Lambda_t} \{\max\{\|\lambda(\cdot) - \text{id}(\cdot)\|, \|x_1(\cdot) - x_2(\lambda(\cdot))\|\}\} \quad (3.45)$$

and

$$d_\infty(x_1, x_2) := \int_0^\infty e^{-t} [d_t(x_1, x_2) \wedge 1] dt. \quad (3.46)$$

It is possible to show (see, e.g., [35]) that this is the correct way of extending the metric, and thus the topology, from $\mathcal{D}([0, t], \mathbb{R})$ to $\mathcal{D}([0, \infty), \mathbb{R})$ in the sense that convergence with respect to d_∞ is equivalent to convergence with respect to d_t on any compact subset of the form $[0, t]$.

The general idea to prove Theorem 6 is to first show the weak convergence of the (rescaled) process $(N_n(k))_{k \geq 0}$ defined in (3.5) and then to deduce the convergence of the reflected process $(\phi(N_n)(k))_{k \geq 0}$ exploiting the following continuous-mapping theorem. In fact, this is a special case of a general technique known as the *continuous-mapping approach* (in [35] and [36] a detailed description of this technique is given). Through the continuous-mapping approach one reduces the problem of establishing convergence of random objects to one of continuity of suitable functions:

Theorem 9 (Continuous Mapping Theorem). *If $X_n \xrightarrow{d} X$ and f is continuous almost surely with respect to the distribution of X , then $f(X_n) \xrightarrow{d} f(X)$.*

Suppose we have shown that $n^{-1/3}N_n(\cdot n^{2/3}) \xrightarrow{d} W(\cdot)$. To prove Theorem 6 we are left to prove that the function $\phi : \mathcal{D} \rightarrow \mathcal{D}$ given by

$$\phi : f(\cdot) \mapsto f(\cdot) - \inf_{y \leq \cdot} f^-(y) \quad (3.47)$$

is continuous almost surely with respect to the distribution of $W(\cdot)$. For this note that $\mathbb{P}(W(\cdot) \in \mathcal{C}) = 1$, where $\mathcal{C} = \mathcal{C}([0, \infty), \mathbb{R}) \subset \mathcal{D}$ denotes the space of continuous functions from $[0, \infty)$ to \mathbb{R} .

Then by [35, Theorem 4.1] and [35, Theorem 6.1], ϕ is continuous almost surely with respect to the distribution of $W(\cdot)$.

To prove the convergence of the rescaled process $(N_n(k))_{k \geq 1}$ we make use of a well-known martingale functional central limit theorem (see [11, Section 7]) in the special case in which the limit process is a standard Brownian motion. For a thorough overview of this important case, see [37]. For completeness we include the theorem that we shall apply, which is the one-dimensional version of the one contained in [37]. Recall that if $M(t)$ is a square-integrable martingale with respect to a filtration $\{\mathcal{F}_t\}_{t \geq 0}$, the *predictable quadratic variation process* associated with $M(t)$ is the unique nondecreasing, nonnegative, predictable, integrable process $V(t)$ such that $M^2(t) - V(t)$ is a martingale with respect to $\{\mathcal{F}_t\}_{t \geq 0}$.

Theorem 10 (MFCLT). *Let $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$ be an increasing filtration and $\{\bar{M}_n\}_{n \in \mathbb{N}}$ be a sequence of continuous-time, real-valued, square-integrable martingales, each with respect to \mathcal{F}_n , such that $\bar{M}_n(0) = 0$. Assume that $(\bar{V}_n(t))_{t \geq 0}$, the predictable quadratic variation process associated with $(\bar{M}_n(t))_{t \geq 0}$, and $\bar{M}_n(t)$ satisfy the following conditions:*

- (a) $\bar{V}_n(t) \xrightarrow{\mathbb{P}} \sigma^2 t, \quad \forall t \in [0, \infty];$
- (b) $\lim_{n \rightarrow \infty} \mathbb{E}[\sup_{t \leq \bar{t}} |\bar{V}_n(t) - \bar{V}_n(t^-)|] = 0, \quad \forall \bar{t} \in \mathbb{R}^+;$
- (c) $\lim_{n \rightarrow \infty} \mathbb{E}[\sup_{t \leq \bar{t}} |\bar{M}_n(t) - \bar{M}_n(t^-)|^2] = 0, \quad \forall \bar{t} \in \mathbb{R}^+.$

Then, as $n \rightarrow \infty$, \bar{M}_n converges in distribution in $\mathcal{D}([0, \infty))$ to a centered Brownian motion with variance $\sigma^2 t$.

In order to apply the MFCLT, we apply a Doob-type decomposition to the process $N_n(k)$ to write it as the sum of a martingale—which will converge to the Brownian motion—and an appropriate drift term:

$$\begin{aligned} N_n(k) &= \sum_{i=1}^k (A_n(i) - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}]) + \sum_{i=1}^k (\mathbb{E}[A_n(i)|\mathcal{F}_{i-1}] - 1) \\ &=: M_n(k) + C_n(k), \end{aligned} \tag{3.48}$$

with $\{\mathcal{F}_i\}_{i \geq 1}$ the filtration generated by $(A_n(k))_{k \geq 1}$, i.e. $\mathcal{F}_i = \sigma(\{A_n(k)\}_{k=1}^i)$. Another Doob decomposition of interest is

$$M_n^2(k) = Z_n(k) + V_n(k) \tag{3.49}$$

with $Z_n(k)$ a martingale and $V_n(k)$ the discrete-time predictable quadratic variation of the process $M_n(k)$. Note that for every fixed n and k , $|M_n(k)|$ is bounded and thus has a finite second moment. Therefore $V_n(k)$ exists and is given by

$$\begin{aligned} V_n(k) &= \sum_{i=1}^k \mathbb{E}[(A_n(i) - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}])^2 | \mathcal{F}_{i-1}] \\ &= \sum_{i=1}^k (\mathbb{E}[A_n(i)^2 | \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}]^2). \end{aligned} \tag{3.50}$$

To see this, we rewrite

$$\begin{aligned}
M_n^2(k) &= \sum_{i=1}^k (A_n(i) - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}])^2 + \sum_{\substack{i,j \leq k \\ i \neq j}} (A_n(i) - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}]) (A_n(j) - \mathbb{E}[A_n(j)|\mathcal{F}_{j-1}]) \\
&=: \sum_{i=1}^k (A_n(i) - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}])^2 + L_n(k).
\end{aligned} \tag{3.51}$$

By developing $L_n(k)$ one can easily see that it is a martingale. The decomposition (3.49) then follows from:

$$\begin{aligned}
Z_n(k) &:= \sum_{i=1}^k (A_n(i) - \mathbb{E}[A(i)|\mathcal{F}_{i-1}])^2 - \sum_{i=1}^k \mathbb{E}[(A_n(i) - \mathbb{E}[A(i)|\mathcal{F}_{i-1}])^2 | \mathcal{F}_{i-1}] + L_n(k), \\
V_n(k) &:= \sum_{i=1}^k \mathbb{E}[(A_n(i) - \mathbb{E}[A(i)|\mathcal{F}_{i-1}])^2 | \mathcal{F}_{i-1}].
\end{aligned} \tag{3.52}$$

Note that $Z_n(k)$ is the sum of two martingales and thus is itself a martingale.

4 Proof of Theorem 5

Note that, when the arrival clocks are exponentially distributed, the conditional law of $A_n(k)$ in (3.4) (conditioned on $\{A_n(k-1), \sum_{j=1}^{k-1} D_k\}$) is distributed as

$$\sum_{i=1}^{P_n(k)} \mathbb{1}_{\{T_{i,k} \leq D_k\}}, \tag{4.1}$$

where $T_{i,k} \stackrel{d}{=} T_i$, which means that the clocks are re-drawn after each service, and $P_n(k) := \lfloor [n] \setminus \nu_k \rfloor = n - Q_n(k-1) - k$ is the number of customers still in the population. Since the dependence of $T_{i,k}$ on k does not play a role in our analysis, we will write T_i instead of $T_{i,k}$. By directly defining (4.1) as the random variable that describes the new arrivals during the k -th service time, this case becomes significantly lighter on the technicalities.

Note also that the exponential distribution satisfies assumptions (3.21) and (3.24). However, Theorem 5 actually holds for all random variables $(T_i)_{i \geq 1}$ satisfying

$$F_T(x) = f_T(0) \cdot x + o(x^{4/3}), \tag{4.2}$$

which is strictly weaker than (3.21).

4.1 Supporting lemmas

We begin with an important lemma, which should be thought of as stating that, for the small-o term in the Taylor expansions we exploit, it holds that

$$\mathbb{E}[o_{\mathbb{P}}(x)] = o(x).$$

Lemma 5. *Let S, T be positive random variables. Define $D := S/n$. If (4.2) holds for T and $\mathbb{E}[S^2] < \infty$, then*

$$\mathbb{E}[\mathbb{P}(T \leq D|D) - f_T(0) \cdot D] = o(n^{-4/3}), \quad (4.3)$$

$$\mathbb{E}[|D(\mathbb{P}(T \leq D|D) - f_T(0) \cdot D)|] = o(n^{-2}), \quad (4.4)$$

$$\mathbb{E}[|\mathbb{P}(T \leq D|D) - f_T(0) \cdot D|^2] = o(n^{-2}). \quad (4.5)$$

Proof. Since

$$\mathbb{P}(T \leq D|D) = F_T(D) = f_T(0) \cdot D + o(S^{4/3}n^{-4/3}), \quad (4.6)$$

pointwise convergence trivially holds:

$$n^{4/3}|F_T(D) - f_T(0) \cdot D| \xrightarrow{\text{a.s.}} 0, \quad \text{as } n \rightarrow \infty. \quad (4.7)$$

Moreover, there exists a constant $C > 0$ such that

$$|F_T(x) - f_T(0) \cdot x| \leq Cx^{4/3}, \quad (4.8)$$

which implies

$$n^{4/3}|F_T(D) - f_T(0) \cdot D| \stackrel{\text{a.s.}}{\leq} n^{4/3} \cdot C(S/n)^{4/3}. \quad (4.9)$$

Since $\mathbb{E}[S^{4/3}] < \infty$, the random variable $n^{4/3}|F_T(D) - f_T(0) \cdot D|$ is bounded by an integrable random variable not depending on n . We can then apply the Dominated Convergence Theorem to obtain (4.3). Using an analogous argument we can also prove (4.4) and (4.5). Pointwise convergence is again trivial. Dominance is obtained by noting that there exists constant $C_1, C_2 > 0$ such that

$$x|F_T(x) - f_T(0) \cdot x| \leq C_1x^2 \quad \text{and} \quad |F_T(x) - f_T(0) \cdot x|^2 \leq C_2x^2. \quad (4.10)$$

Indeed, for $x \ll 1$, $|F_T(x) - f_T(0) \cdot x|^2 \leq C_2x^{8/3} \leq C_2x^2$ for some $N_2 > 0$, and for $x \gg 1$ it is enough to notice that $F_T(x)$ is bounded. The first bound in (4.10) is obtained in the same way. Since $\mathbb{E}[S^2] < \infty$ by assumption, (4.4) and (4.5) again follow by dominated convergence. \square

A random variable that plays a major role in the proof is the one associated with the arrivals when the queue would not deplete,

$$A'_n := \sum_{i=1}^n \mathbf{1}_{\{T_i \leq D\}}. \quad (4.11)$$

Note that

$$A_n(k) \preceq A'_n, \quad \forall k \geq 1. \quad (4.12)$$

Before turning to the proof of Theorem 6 in Section 4, we establish some lemmas that will prove to be useful both in the proof and in providing some insight into the behavior of the process $N_n(\cdot)$.

Lemma 6. *Let $N_n(\cdot)$ be the process defined in (3.5). Then, for $k = O(n^{2/3})$, $n^{-2/3}N_n(k) \preceq G_n(k)$, where $G_n(k)$ is a random variable such that $G_n(k) \xrightarrow{\mathbb{P}} 0$.*

Proof. First note that

$$N_n(k-1) \preceq \sum_{j=1}^{k-1} \left(\sum_{l=1}^n \mathbf{1}_{\{T_l \leq D_j\}} - 1 \right) = \sum_{j=1}^{k-1} (A'_n - 1). \quad (4.13)$$

By the Weak LLN for uncorrelated random variables (see e.g. [25]) it is enough to prove that $\sup_{n \in \mathbb{N}} \text{Var}(A'_n) < \infty$. Write

$$\begin{aligned} \text{Var}(A'_n)^2 &= \mathbb{E}[(A'_n)^2] - \mathbb{E}[A'_n]^2 \\ &= \mathbb{E}[A'_n] + \mathbb{E}\left[\sum_{\substack{i,j \leq n \\ i \neq j}} \mathbf{1}_{\{T_i \leq D\}} \mathbf{1}_{\{T_j \leq D\}}\right] - \mathbb{E}[A'_n]^2. \end{aligned} \quad (4.14)$$

The terms $\mathbb{E}[A'_n]$ and $\mathbb{E}[A'_n]^2$ are uniformly bounded since

$$\mathbb{E}[A'_n] = 1 + \beta n^{-1/3} + o(n^{-1/3}). \quad (4.15)$$

Moreover,

$$\mathbb{E}\left[\sum_{\substack{i,j \leq n \\ i \neq j}} \mathbf{1}_{\{T_i \leq D\}} \mathbf{1}_{\{T_j \leq D\}}\right] = \sum_{\substack{i,j \leq n \\ i \neq j}} \mathbb{E}[F_T(D)^2] \leq C + o(1), \quad (4.16)$$

for some $C > 1$, where we have used Lemma 5 to bound the terms of lower order than D^2 appearing when Taylor expanding $F_T(D)^2$. Both error terms in (4.15) and (4.16) can be bounded from above by a constant independent of n . Therefore $\sup_{n \in \mathbb{N}} \text{Var}(A'_n) < \infty$ and the Weak LLN applies. \square

Intuitively Lemma 6 states that the process $(N_n(k))_{k \geq 1}$ is of order smaller than $n^{2/3}$. Note however that the convergence we established is not uniform in $j \leq k$, with $k = O(n^{2/3})$. We now work towards this result.

We will make use of the following well-known lemma for the order statistics of exponential random variables. Recall that $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ denote the order statistics of random variables X_1, \dots, X_n .

Lemma 7. *Let E_1, \dots, E_n be independent exponentially distributed random variables with mean one. Then,*

$$(E_{(j)})_{j=1}^n \stackrel{d}{=} \left(\sum_{s=1}^j \frac{E_s}{n-s+1} \right)_{j=1}^n.$$

In particular we can couple $(E_{(j)})_{j=1}^n$ with $(E_j)_{j=1}^n$ in such a way that for all $j \leq n$,

$$E_{(j)} \stackrel{\text{a.s.}}{\geq} \frac{E_1 + \dots + E_j}{n}. \quad (4.17)$$

See for example [10, Section 2.5] for a proof of Lemma 7.

We next investigate the random variable A'_n . The following lemma states that on average, in the limit, the contribution to the queue length of arrivals of order $n^{1/3}$ or greater is negligible:

Lemma 8. Define A'_n as in (4.11). Then $A_n'^2$ is stochastically dominated by a family of uniformly integrable (with respect to n) random variables. In particular,

$$\mathbb{E}[(A'_n)^2 \mathbb{1}_{\{(A'_n)^2 > \varepsilon n^{2/3}\}}] \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (4.18)$$

Proof. Observe that, if $E_i \sim \text{Exp}(1)$, $u_i = 1 - \exp(-E_i)$ is a uniform random variable on $[0, 1]$ so that

$$A'_n = \sum_{i=1}^n \mathbb{1}_{\{T_i \leq D\}} = \sum_{i=1}^n \mathbb{1}_{\{T_{(i)} \leq D\}} \stackrel{d}{=} \sum_{i=1}^n \mathbb{1}_{\{F_T^{-1}(1 - \exp(-E_{(i)})) \leq D\}}. \quad (4.19)$$

Since the function $x \mapsto F_T^{-1}(1 - \exp(-x))$ is monotone, by using the coupling in Lemma 7,

$$\begin{aligned} \sum_{i=1}^n \mathbb{1}_{\{F_T^{-1}(1 - \exp(-E_{(i)})) \leq D\}} &\preceq \sum_{i=1}^n \mathbb{1}_{\{F_T^{-1}(1 - \exp(-\sum_{j=1}^i E_j/n)) \leq D\}} \\ &\stackrel{\text{a.s.}}{=} \sum_{i=1}^n \mathbb{1}_{\{\sum_{j=1}^i E_j \leq -n \log(1 - F_T(D))\}}. \end{aligned} \quad (4.20)$$

By (4.2), $F_T(x)/x$ is bounded from above by a positive constant $K \in \mathbb{R}^+$, so that

$$\sum_{i=1}^n \mathbb{1}_{\{\sum_{j=1}^i E_j \leq -n \log(1 - F_T(D))\}} \preceq \sum_{i=1}^n \mathbb{1}_{\{\sum_{j=1}^i E_j \leq -n \log(1 - K \cdot D)\}}. \quad (4.21)$$

Fix ε and find the respective C such that $-\log(1 - x) \leq Cx$ for all $0 \leq x \leq 1 - \varepsilon$. We do this in order to remove the dependencies from n . We then obtain that

$$\begin{aligned} A'_n &\preceq \left(\sum_{i=1}^{\infty} \mathbb{1}_{\{\sum_{j=1}^i E_j \leq CK \cdot nD\}} \right) \mathbb{1}_{\{K \cdot D \leq 1 - \varepsilon\}} + A'_n \mathbb{1}_{\{K \cdot D > 1 - \varepsilon\}} \\ &\preceq N(CK \cdot nD) + A'_n \mathbb{1}_{\{K \cdot D > 1 - \varepsilon\}}, \end{aligned} \quad (4.22)$$

where

$$N(t, \omega) := \sum_{i=1}^{\infty} \mathbb{1}_{\{\sum_{j=1}^i E_j \leq t\}}(\omega) \quad (4.23)$$

is a Poisson process with rate one. We now prove that each of the two terms in (4.22) is a family of uniformly integrable random variables, and thus also their sum is. Since by assumption S_k has a finite second moment, also $N(CK \cdot nD)$ has. Since the latter does not depend on n , it is uniformly integrable with respect to n . Moving to the second term, note that, since $A'_n \leq n$ a.s.,

$$\begin{aligned} \mathbb{E}[A_n'^2 \mathbb{1}_{\{K \cdot D > 1 - \varepsilon\}}] &\leq n^2 \mathbb{P}\left(S_k > \frac{(1 - \varepsilon)n}{K(1 + \beta n^{-1/3})}\right) \\ &\leq n^2 K^2 (1 + \beta n^{-1/3})^2 \frac{\mathbb{E}[S_k^2 \mathbb{1}_{\{S_k > \frac{(1 - \varepsilon)n}{K(1 + \beta n^{-1/3})}\}}]}{(1 - \varepsilon)^2 n^2}. \end{aligned} \quad (4.24)$$

Since $\mathbb{E}[S^2] < \infty$,

$$\mathbb{E}[S_k^2 \mathbf{1}_{\{S_k > \frac{(1-\varepsilon)n}{K(1+\beta n^{-1/3})}\}}] \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (4.25)$$

The second moments of the second term in (4.22) converge to zero as n goes to infinity, and thus $\{A_n'^2 \mathbf{1}_{\{K \cdot D > 1-\varepsilon\}}\}_{n \geq 1}$ is a uniformly integrable family. Therefore, $(N(CK \cdot nD) + A_n' \mathbf{1}_{\{K \cdot D > 1-\varepsilon\}})^2$ is uniformly integrable. We have then shown that $(A_n')_{n \geq 1}$ is stochastically dominated by a random variable with uniformly integrable second moments. The second claim then follows by the stochastic domination result in (3.44). \square

4.2 Proof of Theorem 5

Recall that $N_n(k)$ can be decomposed as $N_n(k) = M_n(k) + C_n(k)$, where $M_n(k)$ is a martingale and $C_n(k)$ is the drift term. Moreover, $M_n^2(k)$ was also written as $M_n^2(k) = Z_n(k) + V_n(k)$ with $Z_n(k)$ the Doob martingale and $V_n(k)$ its drift. The proof then consists of verifying the following conditions:

- (i) $\sup_{t \leq \bar{t}} |n^{-1/3} C_n(tn^{2/3}) - \beta t + \frac{1}{2} t^2| \xrightarrow{\mathbb{P}} 0, \quad \forall \bar{t} \in \mathbb{R}^+;$
- (ii) $n^{-2/3} V_n(tn^{2/3}) \xrightarrow{\mathbb{P}} \sigma^2 t, \quad \forall t \in \mathbb{R}^+;$
- (iii) $\lim_{n \rightarrow \infty} n^{-2/3} \mathbb{E}[\sup_{t \leq \bar{t}} |V_n(tn^{2/3}) - V_n(tn^{2/3}-)|] = 0, \quad \forall \bar{t} \in \mathbb{R}^+;$
- (iv) $\lim_{n \rightarrow \infty} n^{-2/3} \mathbb{E}[\sup_{t \leq \bar{t}} |M_n(tn^{2/3}) - M_n(tn^{2/3}-)|^2] = 0, \quad \forall \bar{t} \in \mathbb{R}^+.$

Recall that $\sigma^2 := f_T^2(0) \mathbb{E}[S_1^2]$.

Condition (i) implies the convergence of the drift term, while conditions (ii)-(iv) imply the convergence of the (rescaled) process $M_n(k)$ to a centered Brownian motion, by Theorem 10. By standard convergence results one can then conclude that the rescaled version of the sum $C_n(k) + M_n(k)$ converges in distribution to the sum of the respective limits.

4.2.1 Proof of (i)

We first prove (i) and to that end, we expand the term $\mathbb{E}[A_n(i) | \mathcal{F}_{i-1}]$. Recall that we have defined ν_i as the set of the customers that are no longer in the population at the beginning of the service of the i -th customer. Then,

$$\begin{aligned} \mathbb{E}[A_n(i) | \mathcal{F}_{i-1}] &= \sum_{s \notin \nu_i} \mathbb{E}[\mathbf{1}_{\{T_s \leq D_i\}} | \mathcal{F}_{i-1}] \\ &= \sum_{s \notin \nu_i} \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{T_s \leq D_i\}} | \mathcal{F}_{i-1}, D_i] | \mathcal{F}_{i-1}] \\ &= \sum_{s \notin \nu_i} \left(\mathbb{E}[f_T(0) D_i | \mathcal{F}_{i-1}] + o(n^{-4/3}) \right), \end{aligned} \quad (4.26)$$

where, in the last equality, we have used Lemma 5. Since D_i is independent from \mathcal{F}_{i-1} , we obtain

$$\mathbb{E}[A_n(i) | \mathcal{F}_{i-1}] = \sum_{s \notin \nu_i} \left(\mathbb{E}[f_T(0) D_i] + o(n^{-4/3}) \right). \quad (4.27)$$

In what follows, we denote by $|A|$ the cardinality of a set A . The summation can then be simplified to

$$\begin{aligned}\mathbb{E}[A_n(i)|\mathcal{F}_{i-1}] &= (n - |\nu_i|) \left(f_T(0) \mathbb{E}[D_i] + o(n^{-4/3}) \right) \\ &= f_T(0) \cdot \mathbb{E}[S_i] (1 + \beta n^{-1/3}) - f_T(0) \mathbb{E}[S_i] (1 + \beta n^{-1/3}) \frac{|\nu_i|}{n} \\ &\quad + (n - |\nu_i|) o(n^{-4/3}),\end{aligned}\tag{4.28}$$

where we note that $|\nu_i|$ is an integer-valued random variable. Then,

$$\mathbb{E}[A_n(i) - 1|\mathcal{F}_{i-1}] = f_T(0)(1 + \beta n^{-1/3}) \mathbb{E}[S_i] - 1 - f_T(0)(1 + \beta n^{-1/3}) \mathbb{E}[S_i] \frac{|\nu_i|}{n} + o(n^{-1/3}),\tag{4.29}$$

By (3.26), $f_T(0)(1 + \beta n^{-1/3}) \mathbb{E}[S_i] = 1 + \beta n^{-1/3} + o(n^{-1/3})$, so that (4.29) yields

$$\mathbb{E}[A_n(i) - 1|\mathcal{F}_{i-1}] = \beta n^{-1/3} - \frac{|\nu_i|}{n} (1 + O(n^{-1/3})) + o(n^{-1/3}).\tag{4.30}$$

Note that the service times are independent from the history of the system, thus the conditioning has no effect. Since $|\nu_i| = i + N_n(i - 1)$, the drift term in the decomposition of $N_n(k)$ can now be written as

$$C_n(k) = k\beta n^{-1/3} - \left(\frac{k^2 + k}{2} + \sum_{i=1}^k N_n(i - 1) \right) \left(\frac{1}{n} + O(n^{-4/3}) \right) + ko(n^{-1/3}).\tag{4.31}$$

The term $-\sum_{i=1}^k N_n(i - 1)$ in (4.31) accounts for the fact that the customers already in the queue cannot rejoin it. This term converges to zero as n tends to infinity (after appropriate scaling) by the following result. In fact, Lemma 9 proves that the process $n^{-2/3}N_n(j)$ tends to zero in probability, uniformly on $j \leq an^{2/3}$, a significantly stronger statement than Lemma 6:

Lemma 9. *Let $(N_n(k))_{k \geq 0}$ be the process defined in (3.5) and $a \in \mathbb{R}^+$. Then*

$$n^{-2/3} \sup_{j \leq an^{2/3}} |N_n(j)| \xrightarrow{\mathbb{P}} 0.\tag{4.32}$$

Proof. Recall that $N_n(j) = M_n(j) + C_n(j)$. Then

$$\begin{aligned}\mathbb{P}(n^{-2/3} \sup_{j \leq an^{2/3}} |N_n(j)| \geq 2\varepsilon) &\leq \mathbb{P}(n^{-2/3} \sup_{j \leq an^{2/3}} |M_n(j)| \geq \varepsilon) \\ &\quad + \mathbb{P}(n^{-2/3} \sup_{j \leq an^{2/3}} |C_n(j)| \geq \varepsilon).\end{aligned}\tag{4.33}$$

We can separately bound the first and second terms. Applying Doob's inequality to the martingale $M_n(\cdot)$ gives

$$\mathbb{P}(n^{-2/3} \sup_{j \leq an^{2/3}} |M_n(j)| \geq \varepsilon) \leq \frac{\mathbb{E}[M_n^2(an^{2/3})]}{(\varepsilon n^{2/3})^2}.\tag{4.34}$$

By (3.49), $\mathbb{E}[M_n^2(k)] = \mathbb{E}[V_n(k)]$. Expanding this term gives

$$\begin{aligned}\mathbb{E}[V_n(k)] &= \mathbb{E} \left[\sum_{i=1}^k (\mathbb{E}[A_n(i)^2|\mathcal{F}_{i-1}] - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}]^2) \right] \\ &\leq \mathbb{E} \left[\sum_{i=1}^k \mathbb{E}[A_n(i)^2|\mathcal{F}_{i-1}] \right] \leq k \cdot \mathbb{E}[A_n'^2],\end{aligned}\tag{4.35}$$

where A'_n is defined as in (4.11). By rescaling we get

$$n^{-4/3}\mathbb{E}[V_n(an^{2/3})] \leq an^{-2/3}\mathbb{E}[A_n'^2]. \quad (4.36)$$

Since A'_n has a finite second moment as was shown in (4.16), the right-most term in (4.36) tends to zero as n tends to infinity.

For the second term in (4.33) we make use of the decomposition of the drift term in (4.31). From there we obtain

$$-(n^{-1} + O(n^{-4/3})) \sum_{i=1}^k \left(i + N_n(i-1) \right) + ko(n^{-1/3}) \leq C_n(k) \leq k\beta n^{-1/3} + ko(n^{-1/3}), \quad (4.37)$$

and thus,

$$\begin{aligned} \sup_{j \leq an^{2/3}} |C_n(j)| &\stackrel{\text{a.s.}}{\leq} (an^{2/3}\beta n^{-1/3} + an^{2/3}o(n^{-1/3})) \\ &\vee \left((n^{-1} + O(n^{-4/3})) \sum_{i=1}^{an^{2/3}} \left(i + N_n(i-1) \right) + an^{2/3}o(n^{-1/3}) \right), \end{aligned} \quad (4.38)$$

since the two sides of the bound (4.37) are monotone functions of k . By rescaling the first term by $n^{-2/3}$ we obtain $a\beta n^{-1/3} + a \cdot o(n^{-1/3})$, which tends to zero almost surely as n goes to infinity. The second term in (4.38) needs more attention. Notice that the function $i \mapsto i + N_n(i)$ is nonnegative and nondecreasing. Thus, isolating the sum, we can bound it through its final term:

$$\sum_{i=1}^{an^{2/3}} \left(i + N_n(i-1) \right) \stackrel{\text{a.s.}}{\leq} an^{2/3}(an^{2/3} + N_n(an^{2/3} - 1)) \quad (4.39)$$

Rescaling by $n^{-2/3}$ we get for the second term in (4.38)

$$\begin{aligned} &n^{-2/3} \left((n^{-1} + O(n^{-4/3})) \sum_{i=1}^{an^{2/3}} \left(i + N_n(i-1) \right) + an^{2/3}o(n^{-1/3}) \right) \\ &\stackrel{\text{a.s.}}{\leq} a(n^{-1} + O(n^{-4/3}))(an^{2/3} + N_n(an^{2/3} - 1)) + a \cdot o(n^{-1/3}) \end{aligned} \quad (4.40)$$

The $O(n^{-4/3})$ term is of lower order than n^{-1} and can be ignored. By Lemma 6, the right-hand side of (4.40) tends to zero in probability as n tends to infinity and this concludes the proof that the second term in (4.33) converges in probability to zero. \square

Substituting $k = tn^{2/3}$ into (4.31) and multiplying by $n^{-1/3}$ yields

$$n^{-1/3}C_n(tn^{2/3}) = \beta t - \left(\frac{t^2 + tn^{-2/3}}{2} + \sum_{i=1}^k N_n(i-1) \right) \left(1 + O(n^{-1/3}) \right) + o(1). \quad (4.41)$$

Both the small- o and the big- O terms in (4.41) are independent of t (or, equivalently, of k). Indeed, the small- o term originates from Lemma 5 (and is therefore independent of k) and the big- O term was introduced in (4.30) and depends only on n and β . Therefore, the convergence of $n^{-1/3}C_n(tn^{2/3})$ is uniform in $t \leq \bar{t}$ for fixed \bar{t} as required, and this concludes the proof of Lemma 9 and thus of (i).

4.2.2 Proof of (ii)

In order to prove (ii) we first compute

$$\begin{aligned}\mathbb{E}[A_n(i)^2|\mathcal{F}_{i-1}] &= \mathbb{E}\left[\sum_{j \notin \nu_i} \mathbb{1}_{\{T_j \leq D_i\}}^2 + \sum_{\substack{l \neq m \\ l, m \notin \nu_i}} \mathbb{1}_{\{T_l \leq D_i\}} \mathbb{1}_{\{T_m \leq D_i\}} \middle| \mathcal{F}_{i-1}\right] \\ &= \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}] + \mathbb{E}\left[\sum_{\substack{l \neq m \\ l, m \notin \nu_i}} \mathbb{1}_{\{T_l \leq D_i\}} \mathbb{1}_{\{T_m \leq D_i\}} \middle| \mathcal{F}_{i-1}\right],\end{aligned}\quad (4.42)$$

which yields

$$\begin{aligned}\mathbb{E}[A_n(i)^2|\mathcal{F}_{i-1}] - \mathbb{E}[A_n(i)|\mathcal{F}_{i-1}] &= \sum_{\substack{l \neq m \\ l, m \notin \nu_i}} \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{T_l \leq D_i\}} \mathbb{1}_{\{T_m \leq D_i\}} | \mathcal{F}_{i-1}, D_i] | \mathcal{F}_{i-1}] \\ &= \sum_{\substack{l \neq m \\ l, m \notin \nu_i}} \mathbb{E}[\mathbb{P}(T_l \leq D_i | D_i) \mathbb{P}(T_m \leq D_i | D_i) | \mathcal{F}_{i-1}] \\ &= \sum_{\substack{l \neq m \\ l, m \notin \nu_i}} \mathbb{E}[F_T(D_i)^2 | \mathcal{F}_{i-1}].\end{aligned}\quad (4.43)$$

Using (4.2) we can rewrite the summation term as

$$\mathbb{E}[F_T(D_i)^2 | \mathcal{F}_{i-1}] = f_T(0)^2 \mathbb{E}[D_i^2] + 2f_T(0) \mathbb{E}[D_i(F_T(D_i) - f_T(0)D_i)] + \mathbb{E}[(F_T(D_i) - f_T(0)D_i)^2]. \quad (4.44)$$

The second and third terms are $o(n^{-2})$ by Lemma 5. Therefore, we rewrite (4.43) as

$$\sum_{\substack{l \neq m \\ l, m \notin \nu_i}} \left(\frac{f_T(0)^2}{n^2} (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] + o(n^{-2}) \right) = \frac{|\Xi_i| f_T(0)^2}{n^2} (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] + o(1). \quad (4.45)$$

Note that the cardinality of the set $\Xi_i := \{(l, m) : l \neq m, l, m \notin \nu_i\}$ is

$$|\Xi_i| = (n - N_n(i-1) - i)^2 - (n - N_n(i-1) - i), \quad (4.46)$$

thus of the order n^2 . Then, for $k = O(n^{2/3})$,

$$\begin{aligned}C_n(k) &= \sum_{i=1}^k (\mathbb{E}[A_n(i)^2 | \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) | \mathcal{F}_{i-1}]^2) \\ &= \sum_{i=1}^k \left(\frac{|\Xi_i| f_T(0)^2}{n^2} (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] + \mathbb{E}[A_n(i) | \mathcal{F}_{i-1}] - \mathbb{E}[A_n(i) | \mathcal{F}_{i-1}]^2 + o(1) \right).\end{aligned}\quad (4.47)$$

Using (4.30), together with the observation that $|\nu_i|/n = O_{\mathbb{P}}(n^{-1/3})$ uniformly for $i = O(n^{2/3})$,

$$\begin{aligned}C_n(k) &= \sum_{i=1}^k \left(\frac{|\Xi_i| f_T(0)^2}{n^2} (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] + O_{\mathbb{P}}(n^{-1/3}) + o(1) \right) \\ &= \sum_{i=1}^k \frac{(n - N_n(i-1) - i)^2 - (n - N_n(i-1) - i)}{n^2} f_T(0)^2 (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] \\ &\quad + O_{\mathbb{P}}(kn^{-1/3}) + o(k).\end{aligned}\quad (4.48)$$

We then split the term inside the summation to isolate the contribution of the process N_n , and write

$$\begin{aligned} V_n(k) &= \sum_{i=1}^k \frac{(n-i)^2 - (n-i)}{n^2} f_T(0)^2 (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] \\ &\quad + \sum_{i=1}^k \frac{N_n(i-1)(N_n(i-1) - 2(n-i) + 1)}{n^2} f_T(0)^2 (1 + \beta n^{-1/3})^2 \mathbb{E}[S^2] \\ &\quad + O_{\mathbb{P}}(kn^{-1/3}) + o(k). \end{aligned} \quad (4.49)$$

By Lemma 9 the second term (accounting for the process history) tends to zero in probability when rescaled appropriately. An elementary computation shows that

$$\begin{aligned} \frac{f_T(0)^2 \mathbb{E}[S^2]}{n^2} \sum_{l=n-k}^{n-1} (l^2 - l) &= \frac{\sigma^2}{n^2} \left(\frac{2}{3}k + k^2 - \frac{1}{3}k^3 - 2kn - k^2n + kn^2 \right) \\ &= \sigma^2 k + O(k^2 n^{-1}). \end{aligned} \quad (4.50)$$

The remaining terms were omitted because they are of order smaller than $O(k^2 n^{-1})$ when $k = sn^{2/3}$. When rescaling space and time appropriately in (4.49) we finally obtain that

$$n^{-2/3} V_n(tn^{2/3}) \xrightarrow{\mathbb{P}} \sigma^2 t, \quad (4.51)$$

as required. This completes the proof of (ii).

4.2.3 Proof of (iii)

It is well known that $(V_n(k))_{k \in \mathbb{N}}$ in (3.49) is almost surely increasing. For (iii), we need to estimate the largest possible jump

$$n^{-2/3} |V_n(k+1) - V_n(k)| = n^{-2/3} |\mathbb{E}[A_n(k+1)^2 | \mathcal{F}_k] - \mathbb{E}[A_n(k+1) | \mathcal{F}_k]^2|, \quad (4.52)$$

with $k = O(n^{2/3})$. In order to apply the Dominated Convergence Theorem, we first prove that it converges pointwise to zero, and then that it is (almost surely) bounded. The jump (4.52) has already been implicitly computed as the term in the summation in (4.49) and it takes the form

$$\begin{aligned} &n^{-2/3} |\mathbb{E}[A_n(k+1)^2 | \mathcal{F}_k] - \mathbb{E}[A_n(k+1) | \mathcal{F}_k]^2| \\ &= n^{-\frac{2}{3}} \left| f_T(0)^2 \mathbb{E}[S^2] \left(\frac{(n-k-1)^2 - (n-k-1)}{n^2} + \frac{N_n(k)(N_n(k) - 2(n-k-1) + 1)}{n^2} \right) \right. \\ &\quad \left. + O_{\mathbb{P}}(n^{-\frac{1}{3}}) \right| \end{aligned} \quad (4.53)$$

The $O_{\mathbb{P}}(n^{-1/3})$ term is a byproduct of $\mathbb{E}[A_n(k) | \mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]^2$ (as computed in (4.47) and following calculations). We now compute it precisely by using the exact expression for $\mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]$ found in (4.30), as follows

$$|\mathbb{E}[A_n(k) | \mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]^2| \stackrel{\text{a.s.}}{\leq} \beta n^{-1/3} + 2|\nu_k|n^{-1} + 2\beta|\nu_k|n^{-4/3} + o(n^{-1/3}). \quad (4.54)$$

The right-hand side of (4.54) is bounded by 3 for all sufficiently large values of n , uniformly in $k \leq tn^{2/3}$ (recall that $|\nu_k| \leq n$). Then, by inserting this into (4.53),

$$\begin{aligned} & n^{-2/3} |\mathbb{E}[A_n(k+1)^2 | \mathcal{F}_k] - \mathbb{E}[A_n(k+1) | \mathcal{F}_k]^2| \\ & \stackrel{\text{a.s.}}{\leq} n^{-2/3} \left| \sigma^2 \left(\frac{(n-k-1)^2 - (n-k-1)}{n^2} + \frac{N_n(k)(N_n(k) - 2(n-k-1) + 1)}{n^2} \right) + 3 \right|. \end{aligned} \quad (4.55)$$

Since $|N_n(k)| \leq n$, we can find a constant C such that, uniformly in $k \leq tn^{2/3}$,

$$n^{-2/3} |\mathbb{E}[A_n(k+1)^2 | \mathcal{F}_k] - \mathbb{E}[A_n(k+1) | \mathcal{F}_k]^2| \stackrel{\text{a.s.}}{\leq} n^{-2/3} (3 + \sigma^2 C). \quad (4.56)$$

This gives both the assumptions in the Dominated Convergence Theorem (almost sure convergence and dominance with an integrable random variable) and therefore concludes the proof of (iii). \square

4.2.4 Proof of (iv)

We tackle (iv) through a coupling argument. First observe that

$$\begin{aligned} n^{-2/3} \mathbb{E}[\sup_{t \leq \bar{t}} |M_n(tn^{2/3}) - M_n(tn^{2/3-})|^2] &= n^{-2/3} \mathbb{E}[\sup_{k \leq \bar{t}n^{2/3}} |A_n(k) - \mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]|^2] \\ &\leq n^{-2/3} \mathbb{E}[\sup_{k \leq \bar{t}n^{2/3}} |A_n(k)|^2] + n^{-2/3} \mathbb{E}[\sup_{k \leq \bar{t}n^{2/3}} |\mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]|^2]. \end{aligned} \quad (4.57)$$

The second term in (4.57) is easily estimated. Using the calculations on $\mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]$ done in (4.30), and that the second term there is negative, yields

$$0 \stackrel{\text{a.s.}}{\leq} \mathbb{E}[A_n(k) | \mathcal{F}_{k-1}] \stackrel{\text{a.s.}}{\leq} 1 + O(n^{-1/3}).$$

For the first term, we will use a coupling argument. For $\varepsilon > 0$, we split

$$\begin{aligned} \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} |A_n(k)|^2\right] &= \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \mathbf{1}_{\{\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \leq \varepsilon n^{2/3}\}}\right] \\ &\quad + \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \mathbf{1}_{\{\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 > \varepsilon n^{2/3}\}}\right]. \end{aligned} \quad (4.58)$$

After multiplying (4.58) by $n^{-2/3}$, the first term can trivially be bounded by ε , while for the second term we estimate

$$\begin{aligned} \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \mathbf{1}_{\{\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 > \varepsilon n^{2/3}\}}\right] &\leq \sum_{k=1}^{\bar{t}n^{2/3}} \mathbb{E}[A_n(k)^2 \mathbf{1}_{\{A_n(k)^2 > \varepsilon n^{2/3}\}}] \\ &\leq \sum_{k=1}^{\bar{t}n^{2/3}} \mathbb{E}[A_n'(k)^2 \mathbf{1}_{\{A_n'(k)^2 > \varepsilon n^{2/3}\}}] \\ &= \bar{t}n^{2/3} \mathbb{E}[A_n'^2 \mathbf{1}_{\{A_n'^2 > \varepsilon n^{2/3}\}}], \end{aligned} \quad (4.59)$$

where we have used the stochastic domination in (4.11). By Lemma 8, $\mathbb{E}[A_n'^2 \mathbf{1}_{\{A_n'^2 > \varepsilon n^{2/3}\}}] \rightarrow 0$ and thus

$$n^{-2/3} \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \mathbf{1}_{\{\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 > \varepsilon n^{2/3}\}}\right] \rightarrow 0.$$

This concludes the proof of (iv). \square

5 Proof of Theorem 6

We now generalize Theorem 5 in the direction of allowing for non-exponential arrival times. Consider a family of arrival times $(T_i)_{i=1}^n$ with distribution function $F_T(\cdot)$ and density function $f_T(\cdot)$. Denote their order statistics by $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$. As before, let $(D_j)_{j \geq 1}$ be the rescaled service times with $D_j := S_j(1 + \beta n^{-1/3})/n$. Then, at time k , the number of arrivals during every service is defined as

$$A_n(k) := \sum_{i \notin \nu_k} \mathbb{1}_{\{\sum_{j=1}^{k-1} D_j < T_i \leq \sum_{j=1}^k D_j\}}. \quad (5.1)$$

We assume that $F_T(\cdot)$ can be Taylor expanded in a neighborhood of every point, as in (3.21), and that the density $f_T(\cdot)$ can be Taylor expanded in a neighborhood of zero, as in (3.24).

5.1 Supporting lemmas

For readability, throughout this section we will use the notation $\Sigma_j := \sum_{i=1}^j D_i$. In this section we present several useful lemmas that are appropriate adaptations of lemmas in Section 4.1. We provide the proof only when it is substantially different from their counterparts in Section 4.1. It is often useful to bound $A_n(k)$ from above to avoid the difficulties of dealing with the complicated set ν_k . Mimicking (4.11), we define $A'_n(k)$ as

$$A'_n(k) := \sum_{i=1}^n \mathbb{1}_{\{\Sigma_{k-1} \leq T_i \leq \Sigma_k\}}. \quad (5.2)$$

Note in particular that

$$A_n(k) \stackrel{\text{a.s.}}{\leq} A'_n(k). \quad (5.3)$$

Lemma 10. *In the present setting and for $k = O(n^{2/3})$, the following holds:*

$$\mathbb{E}[|F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) \cdot D_k| | \Sigma_{k-1}] = o_{\mathbb{P}}(n^{-4/3}), \quad (5.4)$$

$$\mathbb{E}[|D_k(F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) \cdot D_k)| | \Sigma_{k-1}] = o_{\mathbb{P}}(n^{-2}), \quad (5.5)$$

$$\mathbb{E}[|F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) \cdot D_k|^2 | \Sigma_{k-1}] = o_{\mathbb{P}}(n^{-2}), \quad (5.6)$$

Moreover, all the statements of convergence hold uniformly for $k = O(n^{2/3})$.

Proof. We give the proof for (5.5), the rest can be shown in an analogous way. Note that, by our assumptions on F_T ,

$$\begin{aligned} & \mathbb{E}[n^2 |D_k \cdot (F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) D_k)| | \Sigma_{k-1}] \\ & \leq \sup_{y \leq Cn^{-1/3}} n^2 \mathbb{E}[|F_T(y + D_k) - F_T(y) - f_T(y) D_k| \cdot D_k] \\ & \leq n^2 \mathbb{E}[\sup_{y \leq Cn^{-1/3}} |F_T(y + D_k) - F_T(y) - f_T(y) D_k| \cdot D_k] \end{aligned} \quad (5.7)$$

since with high probability $\Sigma_{k-1} \leq Cn^{-1/3}$ for some $C > 0$. The right term tends to zero by the Dominated Convergence Theorem and assumption (3.22), and this immediately implies (5.5). \square

As has already been seen for the exponential arrivals case, the process $N_n(k)_{k \geq 1}$ is roughly of the order $n^{1/3}$ around time $tn^{2/3}$. The following lemma is the counterpart for general arrivals of Lemma 6 and we prove it in a similar fashion:

Lemma 11. *Let $N_n(k)$ be the process defined in (3.5). Then $n^{-2/3}N_n(k) \preceq G_n(k)$, where $G_n(k)$ is a random variable such that $G_n(k) \xrightarrow{\mathbb{P}} 0$ for $k = O(n^{2/3})$.*

Proof. For simplicity we only consider $k = tn^{2/3}$. Fix an arbitrary $\varepsilon > 0$. Then,

$$\begin{aligned}
& \mathbb{P}\left(n^{-2/3} \sum_{j=1}^{tn^{2/3}} A_n(j) - n^{-2/3}t \geq \varepsilon\right) \\
& \leq \mathbb{P}\left(n^{-2/3} \sum_{j=1}^{tn^{2/3}} A'_n(j) - n^{-2/3}t \geq \varepsilon\right) \\
& \leq \mathbb{P}\left(n^{-2/3} \left| \sum_{j=1}^{tn^{2/3}} A'_n(j) - n^{-2/3}t \right| \geq \varepsilon\right) \\
& = \mathbb{P}\left(n^{-2/3} \left| \sum_{l=1}^n \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}} - n^{-2/3}t \right| \geq \varepsilon\right) \\
& \leq \frac{\mathbb{E}[(\sum_{l=1}^n \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}})^2] \mathbb{1}_{\{|\sum_{l=1}^n \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}} - n^{-2/3}t| \geq \varepsilon n^{2/3}\}}]}{n^{4/3}\varepsilon^2}. \tag{5.8}
\end{aligned}$$

We are left to bound the expected value in (5.8). To do so, we define the event $\mathcal{E}_n := \{|\sum_{l=1}^n \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}} - n^{-2/3}t| \geq \varepsilon n^{2/3}\}$ and write

$$\begin{aligned}
\mathbb{E}[(\sum_{l=1}^n \mathbb{1}_{\{T_l \leq \Sigma_{tn^{2/3}}\}})^2] \mathbb{1}_{\mathcal{E}_n} & \leq n + \mathbb{E}[\sum_{h \neq k} \mathbb{1}_{\{T_h \leq \Sigma_{tn^{2/3}}\}} \mathbb{1}_{\{T_k \leq \Sigma_{tn^{2/3}}\}} \mathbb{1}_{\mathcal{E}_n}] \\
& \leq n + n^2 \mathbb{E}[F_T(\Sigma_{tn^{2/3}})^2] \mathbb{1}_{\mathcal{E}_n} \leq n + Cn^2 \mathbb{E}[(\Sigma_{tn^{2/3}})^2] \mathbb{1}_{\mathcal{E}_n} \\
& \leq n + Cn^{4/3} \mathbb{E}[S^2] \mathbb{1}_{\mathcal{E}_n}, \tag{5.9}
\end{aligned}$$

for a large constant C . The last inequality in (5.9) was obtained by an application of the Cauchy-Schwarz inequality. Since $\mathbb{P}(\mathcal{E}_n) \rightarrow 0$ and $\mathbb{E}[S^2] < \infty$, by inserting (5.9) into (5.8) and the Dominated Convergence Theorem, we get the desired convergence. \square

The following lemma reduces the task to estimating quantities involving order statistics of exponentials, which are possibly complicated objects, to the one of dealing with a Poisson process, which is much simpler. It will be used later for proving Lemma 13, which is the equivalent in this setting of Lemma 8.

Lemma 12. *Consider the order statistics $(E_{(i)})_{i=1}^{n-v}$ ($v \leq n$) of n exponential unit mean random variables. Define $|\Upsilon_{(0,c)}^{(n-v)}|$ as the cardinality of the set $\Upsilon_{(0,c)}^{(n-v)} := \{j \in [n-v] : E_{(j)} \in (0, c)/n\}$. Then,*

$$|\Upsilon_{(0,c)}^{(n-v)}| \preceq N\left(\frac{n-v}{n}c\right), \tag{5.10}$$

where $N(t)$ is a Poisson process with unit rate.

Proof. The statement is a consequence of Lemma 7. Fix $j \in \{1, \dots, n - v\}$. By definition of stochastic domination

$$\mathbb{P}(E_{(j)} \leq c/n) \leq \mathbb{P}\left(\frac{\sum_{i=1}^j E_i}{n-v} \leq \frac{c}{n}\right) \leq \mathbb{P}\left(\Pi_j \leq \frac{n-v}{n}c\right), \quad (5.11)$$

where Π_j is the j -th point of a Poisson process with rate one. The computation in (5.11) intuitively means that there are more Poisson points in an interval of length $(n-v)\frac{c}{n}$ than order statistics in an interval of length $\frac{c}{n}$. This implies (5.10). \square

Since

$$N\left(\frac{n-v}{n} \cdot c\right) \preceq N(c), \quad \forall v \leq n,$$

it follows from (5.10) that

$$|\Upsilon_{(0,c)}^{(n-v)}| \preceq N(c), \quad \forall v \leq n. \quad (5.12)$$

Corollary 1. *Under the same assumptions as in Lemma 12,*

$$\Upsilon_{(a,b)}^{(n)} \preceq N(b-a). \quad (5.13)$$

Proof. By Lemma 12,

$$\mathbb{P}(N(b-a) \leq x) \leq \mathbb{P}(|\Upsilon_{(0,b-a)}^{(n-v)}| \leq x). \quad (5.14)$$

Note that, by the memoryless property,

$$\mathbb{P}(|\Upsilon_{(0,b-a)}^{(n-v)}| \leq x) \stackrel{\text{a.s.}}{=} \mathbb{P}(|\Upsilon_{(a,b)}^{(n)}| \leq x \mid |\Upsilon_{(0,a)}^{(n)}| = v). \quad (5.15)$$

Since the left side of (5.14) does not depend on v , by combining (5.14) and (5.15) and taking expectations on both sides in order to remove the conditioning, we get

$$\mathbb{P}(N(b-a) \leq x) \leq \mathbb{P}(|\Upsilon_{(a,b)}^{(n)}| \leq x), \quad (5.16)$$

which is (5.13). \square

One of the cornerstones of the analysis in Section 4 was the uniform integrability of $(A_n^{\prime 2})_{n \geq 1}$ in Lemma 8. An analogous version holds in this general setting:

Lemma 13. *$A_n(k)$ is stochastically bounded by a random variable with uniformly integrable (with respect to n) second moment, uniformly in $k \leq tn^{2/3}$.*

Proof. Note that $T_{(i)} \stackrel{d}{=} F_T^{-1}(1 - \exp(-E_{(i)}))$, where $(E_{(i)})_{i=1}^n$ are the order statistics of unit mean exponential random variables. Then,

$$\begin{aligned} A_n(k) &\stackrel{\text{a.s.}}{\leq} A'_n(k) \stackrel{d}{=} \sum_{i=1}^n \mathbb{1}_{\{\Sigma_{k-1} \leq F_T^{-1}(1 - \exp(-E_{(i)})) \leq \Sigma_k\}} \\ &\stackrel{d}{=} \sum_{i=1}^n \mathbb{1}_{\{F_T(\Sigma_{k-1}) - 1 \leq -\exp(-E_{(i)}) \leq F_T(\Sigma_k) - 1\}} \\ &\stackrel{d}{=} \sum_{i=1}^n \mathbb{1}_{\{-\log(1 - F_T(\Sigma_{k-1})) \leq E_{(i)} \leq -\log(1 - F_T(\Sigma_k))\}}. \end{aligned} \quad (5.17)$$

By Corollary 1,

$$A_n(k) \preceq N \left(n \log \left(\frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_k)} \right) \right). \quad (5.18)$$

By splitting the event space Ω we can write (5.18) as

$$A_n(k) \preceq N \left(n \log \left(\frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_k)} \right) \right) \mathbb{1}_{\{\Sigma_k \leq \bar{x}\}} + A_n(k) \mathbb{1}_{\{\Sigma_k > \bar{x}\}}, \quad (5.19)$$

where \bar{x} is independent of n and will be determined later on. The plan now is to show that the first term in (5.19) is bounded by a random variable independent of n and with finite second moment, and that the second term has second moments converging to zero, as n tends to infinity. These two facts together imply that the right-hand side of (5.19) has uniformly integrable second moments.

The first term is easily bounded. We now choose \bar{x} in such a way that $1 - F_T(\bar{x}) > 0$. By Taylor expanding the function $x \mapsto \log \left(\frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_{k-1} + x)} \right)$, we get, for some Σ_{k-1}^* in between Σ_{k-1} and Σ_k

$$\begin{aligned} N \left(n \log \left(\frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_k)} \right) \right) \mathbb{1}_{\{\Sigma_k \leq \bar{x}\}} &= N \left(n \frac{f_T(\Sigma_{k-1}^*)}{1 - F_T(\Sigma_{k-1}^*)} \frac{S_k}{n} \right) \mathbb{1}_{\{\Sigma_k \leq \bar{x}\}} \\ &\preceq N \left(\frac{f_T(0)}{1 - F_T(\bar{x})} S_k \right), \end{aligned} \quad (5.20)$$

where we have used that the density $f_T(\cdot)$ has finite maximum value $f_T(0)$. The right-most term in (5.20) has finite second moment, since $\mathbb{E}[S^2] < \infty$. For the second term we proceed by noting that

$$A_n(k)^2 \mathbb{1}_{\{\Sigma_k \geq \bar{x}\}} \stackrel{\text{a.s.}}{\leq} A_n(k)^2 \mathbb{1}_{\{D_k \geq \bar{x}/2\}} + A_n(k)^2 \mathbb{1}_{\{\Sigma_k \geq \bar{x}, D_k < \bar{x}/2\}}, \quad (5.21)$$

The mean of the first term can be bounded by

$$\mathbb{E}[A_n(k)^2 \mathbb{1}_{\{D_k \geq \bar{x}/2\}}] \leq n^2 \mathbb{P}(S_k \geq n\bar{x}/2) \leq n^2 \frac{\mathbb{E}[S_k^2 \mathbb{1}_{\{S_k \geq n\bar{x}/2\}}]}{(n\bar{x}/2)^2}, \quad (5.22)$$

and the right-hand side tends to zero as n tends to infinity since $\mathbb{E}[S_k^2] < \infty$. For the second term, some more work is needed. First observe that $\mathbb{1}_{\{\Sigma_k \geq \bar{x}, D_k < \bar{x}/2\}} \leq \mathbb{1}_{\{\Sigma_{k-1} \geq \bar{x}/2\}}$. After dominating as usual $A_n(k)^2$ by n^2 , we compute

$$\begin{aligned} \mathbb{E}[A_n(k)^2 \mathbb{1}_{\{\Sigma_k \geq \bar{x}, D_k < \bar{x}/2\}}] &\leq n^2 \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \leq T_i \leq \Sigma_k\}} \mathbb{1}_{\{\Sigma_{k-1} \geq \bar{x}/2\}} | \Sigma_k]] \\ &= n^2 \mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \geq \bar{x}/2\}} \mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \leq T_i \leq \Sigma_k\}} | \Sigma_k]] \\ &= n^2 \mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \geq \bar{x}/2\}} (F_T(\Sigma_k) - F_T(\Sigma_{k-1}))]. \end{aligned} \quad (5.23)$$

By applying the Mean Value Theorem to $F_T(\cdot)$, we obtain

$$|F_T(\Sigma_k) - F_T(\Sigma_{k-1})| \stackrel{\text{a.s.}}{\leq} f_T(0) D_k, \quad (5.24)$$

where we recall that $f_T(0) = \max_{t \in [0, \infty)} f_T(t)$. Inserting this into (5.23),

$$\mathbb{E}[A_n(k)^2 \mathbb{1}_{\{\Sigma_k \geq \bar{x}, D_k < \bar{x}/2\}}] \leq n M \mathbb{E}[S_k \mathbb{1}_{\{\Sigma_{k-1} \geq \bar{x}/2\}}] = n M \mathbb{E}[S_k] \mathbb{P}(\Sigma_{k-1} \geq \bar{x}/2), \quad (5.25)$$

the equality following from independence of S_k and Σ_{k-1} .

It is easy to see that the right-hand side converges to zero by using Chebyshev's inequality. Indeed, taking n so large that $n^{2/3}\mathbb{E}[S] \leq \frac{n\bar{x}}{4}$,

$$\begin{aligned}\mathbb{P}(\Sigma_{k-1} \geq \bar{x}/2) &\leq \mathbb{P}\left(\left|\sum_{i=1}^{tn^{2/3}} S_i - n^{2/3}\mathbb{E}[S_i]\right| \geq n\bar{x}/4\right) \\ &\leq 16 \frac{tn^{2/3}\text{Var}(S_i)}{n^2\bar{x}} = o(n^{-1}).\end{aligned}\tag{5.26}$$

This concludes the proof that the second moment of the second term in (5.19) tends to zero as n tends to infinity. \square

We conclude with a useful application of Doob's inequality:

Lemma 14. *Assume $(S_i)_{i \geq 0}$ is a sequence of i.i.d. random variables with finite second moment. Then, for any $\alpha, \beta > 0$ such that $\alpha < 2\beta$,*

$$\frac{\sup_{k \leq tn^\alpha} \left| \sum_{i=1}^k S_i - k \cdot \mathbb{E}[S] \right|}{n^\beta} \xrightarrow{\mathbb{P}} 0.\tag{5.27}$$

Proof. Define $M_k := \sum_{j=1}^k (S_j - \mathbb{E}[S])$. Then $(M_k)_{k \geq 1}$ is a martingale. Therefore, by Doob's inequality applied to the sub-martingale $(|M_k|)_{k \geq 1}$, we have

$$\mathbb{P}\left(\frac{\sup_{k \leq tn^\alpha} |M_k|}{n^\beta} > \varepsilon\right) \leq \frac{\mathbb{E}[M_{tn^\alpha}^2]}{\varepsilon^2 n^{2\beta}} = \frac{tn^\alpha \mathbb{E}[(S - \mathbb{E}[S])^2]}{\varepsilon^2 n^{2\beta}}.\tag{5.28}$$

This converges to zero since $\alpha < 2\beta$. Note that ε can depend on n , for example by defining $\varepsilon := n^{-\delta}$ and choosing δ such that $\delta < \beta - \alpha/2$. \square

5.2 Proof of Theorem 6

The proof consists, again, in verifying four conditions in order to apply Theorem 10. The first is the convergence of the drift

$$(i) \sup_{t \leq \bar{t}} |n^{-1/3} C_n(tn^{2/3}) - \beta t - f'_T(0)/f_T(0)^2 \cdot \frac{t^2}{2}| \xrightarrow{\mathbb{P}} 0, \quad \forall \bar{t} \in \mathbb{R}^+;$$

while (ii), (iii) and (iv) are technical conditions and analogous to the ones given in Section 4. The filtration we consider henceforth is defined as

$$\mathcal{F}_i := \sigma(\{A_n(j), D_j\}_{j \leq i}).\tag{5.29}$$

5.2.1 Proof of (i)

Computing the asymptotic drift essentially amounts to computing the discrete drift, which in turn depends heavily on

$$\begin{aligned}\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] &= \sum_{i \notin \nu_k} \mathbb{E}\left[\mathbb{1}_{\{\Sigma_{k-1} \leq T_i \leq \Sigma_k\}} \middle| \mathcal{F}_{k-1}\right] \\ &= \sum_{i \notin \nu_k} \mathbb{E}\left[\mathbb{1}_{\{T_i \leq \Sigma_k\}} \middle| \Sigma_{k-1}, \{T_i \geq \Sigma_{k-1}\}\right],\end{aligned}\tag{5.30}$$

where, as above, ν_i denotes the set of the customers that no longer remain in the population at the beginning of the service of the i -th customer. Adding the conditioning on $\{T_i \geq \Sigma_{k-1}\}$ does not influence the conditional expectation, since T_i is such that $i \notin \nu_{k-1}$. Indeed, note that $i \notin \nu_k$ implies $T_i \geq \Sigma_{k-1}$. Then, defining for simplicity $\mathcal{E}_{k-1} := \{\Sigma_{k-1}, \{T_i \geq \Sigma_{k-1}\}\}$,

$$\begin{aligned}\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] &= \sum_{i \notin \nu_k} \mathbb{E}\left[\mathbb{E}[\mathbb{1}_{\{\Sigma_{k-1} \leq T_i \leq \Sigma_k\}} | D_k, \mathcal{E}_{k-1}] | \mathcal{E}_{k-1}\right] \\ &= (n - |\nu_k|) \mathbb{E}\left[\frac{F_T(\Sigma_k) - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_{k-1})} | \mathcal{E}_{k-1}\right] \\ &= \frac{(n - |\nu_k|)}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[F_T(\Sigma_k) - F_T(\Sigma_{k-1}) | \mathcal{E}_{k-1}].\end{aligned}\quad (5.31)$$

We now rearrange the terms in order to distinguish between the ones contributing to the limit and those vanishing, as follows

$$\begin{aligned}\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] - 1 &= \frac{(n - |\nu_k|)}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[F_T(\Sigma_k) - F_T(\Sigma_{k-1}) | \mathcal{E}_{k-1}] - \frac{1 - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_{k-1})} \\ &= \frac{n}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - 1/n | \mathcal{E}_{k-1}] \\ &\quad - \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[|\nu_k|(F_T(\Sigma_k) - F_T(\Sigma_{k-1})) - F_T(\Sigma_{k-1}) | \mathcal{E}_{k-1}] \\ &=: A_n^{(1)}(k) - A_n^{(2)}(k).\end{aligned}\quad (5.32)$$

$A_n^{(1)}(k)$ groups all the terms appearing in the limit, while $A_n^{(2)}(k)$ groups all the terms of lower order, vanishing in the limit. We treat them separately, starting with $A_n^{(1)}(k)$. The term $F_T(\Sigma_k) - F_T(\Sigma_{k-1})$ can be simplified by our assumptions. By (3.21) and Lemma 10,

$$A_n^{(1)}(k) = \frac{n}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[f_T(\Sigma_{k-1})D_k - n^{-1} + o_{\mathbb{P}}(n^{-4/3}) | \mathcal{E}_{k-1}], \quad (5.33)$$

and by (3.24),

$$\begin{aligned}\mathbb{E}[f_T(\Sigma_{k-1})D_k | \mathcal{E}_{k-1}] &= (f_T(0) + f'_T(0)\Sigma_{k-1} + o(\Sigma_{k-1})) \mathbb{E}[D] \\ &= f_T(0)\mathbb{E}[D_k] + f'_T(0)\Sigma_{k-1}\mathbb{E}[D_k] + o(\Sigma_{k-1})\mathbb{E}[D_k],\end{aligned}\quad (5.34)$$

where, with a slight abuse of notation, we denoted $|f_T(\Sigma_{k-1}) - f_T(0) - f'_T(0)\Sigma_{k-1}|$ by $o(\Sigma_{k-1})$. Since, for $k = O(n^{2/3})$,

$$n^{1/3}|f_T(\Sigma_{k-1}) - f_T(0) - f'_T(0)\Sigma_{k-1}| \xrightarrow{\text{a.s.}} 0, \quad (5.35)$$

also convergence in probability holds, that is, $o(\Sigma_{k-1}) = o_{\mathbb{P}}(n^{-1/3})$. In particular $o(\Sigma_{k-1}) \cdot \mathbb{E}[D_k] = o_{\mathbb{P}}(n^{-4/3})$. Inserting (5.34) into (5.33) yields

$$\begin{aligned}A_n^{(1)}(k) &= \frac{n}{1 - F_T(\Sigma_{k-1})} \left(f_T(0)\mathbb{E}[D_k] - 1/n + f'_T(0)\mathbb{E}[D_k]\Sigma_{k-1} + o_{\mathbb{P}}(n^{-4/3}) \right) \\ &= \frac{1}{1 - F_T(\Sigma_{k-1})} \left(f_T(0)\mathbb{E}[S_k] - 1 + \frac{f'_T(0)}{n} \sum_{j=1}^{k-1} S_j \mathbb{E}[S_k] + o_{\mathbb{P}}(n^{-1/3}) \right).\end{aligned}\quad (5.36)$$

The criticality assumption $f_T(0)\mathbb{E}[S_k] = 1 + \beta n^{-1/3} + o(n^{-1/3})$ leads to

$$A_n^{(1)}(k) = \frac{1}{1 - F_T(\Sigma_{k-1})} \left(\beta + f'_T(0)\mathbb{E}[S_k] \frac{\sum_{j=1}^{k-1} S_j}{n^{2/3}} + o_{\mathbb{P}}(1) \right) \cdot n^{-1/3}. \quad (5.37)$$

Recall that the drift term is defined as

$$C_n(s) = \sum_{k=1}^s (\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] - 1) = \sum_{k=1}^s (A_n^{(1)}(k) - A_n^{(2)}(k)) =: C_n^{(1)}(s) + C_n^{(2)}(s). \quad (5.38)$$

We now sum (5.37) over k , obtaining

$$C_n^{(1)}(s) = \sum_{k=1}^s \frac{\beta n^{-1/3}}{1 - F_T(\Sigma_{k-1})} + \frac{f'_T(0)\mathbb{E}[S_1]}{n} \sum_{k=1}^s \frac{\sum_{j=1}^{k-1} S_j}{1 - F_T(\Sigma_{k-1})} + \sum_{k=1}^s \frac{1}{1 - F_T(\Sigma_{k-1})} o_{\mathbb{P}}(n^{-1/3}). \quad (5.39)$$

By scaling time as $s = tn^{2/3}$ and space as $n^{-1/3}$, we obtain

$$n^{-1/3} C_n^{(1)}(tn^{2/3}) = \sum_{k=1}^{tn^{2/3}} \frac{\beta n^{-2/3}}{1 - F_T(\Sigma_{k-1})} + \frac{f'_T(0)\mathbb{E}[S_1]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \frac{\sum_{j=1}^{k-1} S_j}{1 - F_T(\Sigma_{k-1})} + \sum_{k=1}^{tn^{2/3}} \frac{n^{-1/3} o_{\mathbb{P}}(n^{-1/3})}{1 - F_T(\Sigma_{k-1})}. \quad (5.40)$$

Before continuing, we give two easy but useful results:

Lemma 15. *Let $(S_i)_{i \geq 1}$ be a sequence of i.i.d. random variables such that $\mathbb{E}[S_1^2] < \infty$. Then*

$$\left| \frac{\sum_{n=1}^N \sum_{i=1}^n S_i}{N^2} - \frac{\mathbb{E}[S_1]}{2} \right| \xrightarrow{\mathbb{P}} 0, \quad \text{as } N \rightarrow \infty. \quad (5.41)$$

Moreover,

$$\left| \frac{\sum_{n=1}^N \left(\sum_{i=1}^n S_i \right)^2}{N^3} - \frac{\mathbb{E}[S_1]^2}{3} \right| \xrightarrow{\mathbb{P}} 0, \quad \text{as } N \rightarrow \infty. \quad (5.42)$$

Proof. Both claims can be proven through Lemma 14. We omit the details. \square

Another useful ingredient is the following Taylor expansion:

$$\begin{aligned} \frac{1}{1 - F_T(\Sigma_{k-1})} &= 1 + F_T(\Sigma_{k-1}) + \left(\frac{1}{1 - F_T(\Sigma_{k-1})} - 1 - F_T(\Sigma_{k-1}) \right) \\ &= 1 + O_{\mathbb{P}}(\Sigma_{k-1}). \end{aligned} \quad (5.43)$$

In what follows, we compute the limits (in probability) for each term in (5.40).

First term in (5.40). By (5.43) and Lemma 15, we have for the first term

$$\begin{aligned} &\sup_{t \leq \bar{t}} \left| n^{-2/3} \sum_{k=1}^{tn^{2/3}} \frac{1}{1 - F_T(\Sigma_{k-1})} - t \right| \\ &= \sup_{t \leq \bar{t}} \left| n^{-2/3} \sum_{k=1}^{tn^{2/3}} \left(F_T(\Sigma_{k-1}) + \left(\frac{1}{1 - F_T(\Sigma_{k-1})} - 1 - F_T(\Sigma_{k-1}) \right) \right) \right| \\ &\leq n^{-2/3} \sum_{k=1}^{\bar{t}n^{2/3}} 2F_T(\Sigma_{k-1}), \end{aligned}$$

where we have dominated the error term in the Taylor expansion (5.43) by $F_T(\Sigma_{k-1})$ and used the fact that, as a function of t , the summation is an increasing function. We can uniformly dominate F_T as follows

$$\begin{aligned} n^{-2/3} \sum_{k=1}^{\bar{t}n^{2/3}} 2F_T(\Sigma_{k-1}) &\leq n^{-2/3} \sum_{k=1}^{\bar{t}n^{2/3}} 2 \sup_{k \leq \bar{t}n^{2/3}} F_T(\Sigma_{k-1}) \\ &= 2\bar{t}F_T(\Sigma_{\bar{t}n^{2/3}}). \end{aligned} \quad (5.44)$$

The right-hand side of (5.44) tends to zero almost surely, and thus also in probability.

Second term in (5.40). Again, by (5.43), the second term simplifies to

$$\frac{f'_T(0)\mathbb{E}[S]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \frac{\sum_{j=1}^{k-1} S_j}{1 - F_T(\Sigma_{k-1})} = \frac{f'_T(0)\mathbb{E}[S]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \sum_{j=1}^{k-1} S_j + \frac{f'_T(0)\mathbb{E}[S]f_T(0)}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \sum_{j=1}^{k-1} S_j O_{\mathbb{P}}(\Sigma_{k-1}).$$

By the first statement of Lemma 15 the first term converges to $\frac{t^2}{2} \cdot f'_T(0)\mathbb{E}[S_1]^2$ uniformly in t . Indeed,

$$\begin{aligned} \sup_{t \leq \bar{t}} \left| n^{-4/3} \sum_{k=1}^{tn^{2/3}} \sum_{j=1}^{k-1} S_j - \frac{t^2}{2} \mathbb{E}[S] \right| &= \sup_{t \leq \bar{t}} \left| \frac{\sum_{n=1}^{tn^{2/3}} n\mathbb{E}[S]}{n^{4/3}} - \frac{t^2}{2} \mathbb{E}[S] + \frac{\sum_{k=1}^{tn^{2/3}} (\sum_{i=1}^k S_i - k\mathbb{E}[S])}{n^{4/3}} \right| \\ &= \sup_{t \leq \bar{t}} \left| \frac{t\mathbb{E}[S]}{2n^{2/3}} + \frac{\sum_{k=1}^{tn^{2/3}} (\sum_{i=1}^k S_i - k\mathbb{E}[S])}{n^{4/3}} \right| \\ &\stackrel{\text{a.s.}}{\leq} \bar{t} \cdot \frac{\mathbb{E}[S]}{2n^{2/3}} + \bar{t} \cdot \frac{\sup_{k \leq tn^{2/3}} |\sum_{i=1}^k S_i - k\mathbb{E}[S]|}{n^{2/3}}. \end{aligned} \quad (5.45)$$

The second term converges to zero in probability by Lemma 15. Indeed,

$$\sup_{t \leq \bar{t}} \left| \frac{f'_T(0)\mathbb{E}[S]f_T(0)}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \left(\sum_{j=1}^{k-1} S_j \right) O_{\mathbb{P}}(\Sigma_{k-1}) \right| \stackrel{\text{a.s.}}{\leq} \frac{Cf'_T(0)\mathbb{E}[S]f_T(0)}{n^{7/3}} \sum_{k=1}^{\bar{t}n^{2/3}} \left| \left(\sum_{j=1}^{k-1} S_j \right)^2 \right|, \quad (5.46)$$

where we used the domination $O_{\mathbb{P}}(\Sigma_{k-1}) \leq C\Sigma_{k-1}$. The right-most term in (5.46) then converges to zero in probability by Lemma 15.

Third term in (5.40). The error term originates from the Taylor expansion of $n(F_T(\Sigma_k) - F_T(\Sigma_{k-1}))$ done in (5.33). To see that it is uniform in $k \leq \bar{k}$, we write $F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k = \epsilon_k$ and, since by assumption $f'_T(\cdot)$ exists and is continuous in a neighborhood of 0, we bound it as

$$|n\epsilon_k| \leq n \sup_{x \leq Cn^{-1/3}} |f'_T(x)| \frac{D_k^2}{2}, \quad (5.47)$$

similarly as in Lemma 10, when $k = O(n^{2/3})$. In particular,

$$\sup_{k \leq tn^{2/3}} \mathbb{E}[n\epsilon_k | \mathcal{E}_{k-1}] \leq \sup_{x \leq Cn^{-1/3}} |f'_T(x)| \frac{\mathbb{E}[S^2]}{2n} = o(n^{-1/3}), \quad (5.48)$$

and the right-hand side is independent of k . This concludes the bound on the third term in (5.40) and thus the proof that

$$\sup_{t \leq \bar{t}} |n^{-1/3} C_n^{(1)}(tn^{2/3}) - \beta t - t^2 f'_T(0) \mathbb{E}[S]^2 / 2| \xrightarrow{\mathbb{P}} 0. \quad (5.49)$$

□

To conclude, it remains to be proven that $\sup_{t \leq \bar{t}} n^{-1/3} |C_n^{(2)}(tn^{2/3})|$ vanishes in the limit. To do so, we develop the terms of $A_n^{(2)}(k)$ similarly as before, obtaining

$$\begin{aligned} A_n^{(2)}(k) &= \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [|\nu_k| (f_T(\Sigma_{k-1}) D_k + o_{\mathbb{P}}(n^{-1})) - f_T(0) \Sigma_{k-1} + o_{\mathbb{P}}(n^{-1/3}) | \mathcal{E}_{k-1}] \\ &= \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [|\nu_k| f_T(\Sigma_{k-1}) D_k - f_T(0) \Sigma_{k-1} + |\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) | \mathcal{E}_{k-1}] \\ &= \frac{\mathbb{E} [|\nu_k| f_T(0) D_k - f_T(0) \Sigma_{k-1} + f'_T(0) |\nu_k| \Sigma_{k-1} D_k + |\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) | \mathcal{E}_{k-1}]}{1 - F_T(\Sigma_{k-1})}, \end{aligned} \quad (5.50)$$

where $o_{\mathbb{P}}(n^{-1})$ is a convenient notation for $F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) D_k$ and $o_{\mathbb{P}}(n^{-1/3})$ for $F_T(\Sigma_{k-1}) - f_T(0) \Sigma_{k-1}$. We now sum (5.50) over k to obtain

$$\begin{aligned} C_n^{(2)}(s) &= \sum_{k=1}^s \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [|\nu_k| f_T(0) D_k - f_T(0) \Sigma_{k-1} + f'_T(0) |\nu_{k-1}| \Sigma_{k-1} D_k \\ &\quad + |\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) | \mathcal{E}_{k-1}]. \end{aligned} \quad (5.51)$$

Recall that, by definition, $|\nu_k| = k + N_n(k-1)$. Intuitively, k is of a much larger order than $N_n(k-1)$, therefore at first approximation we ignore $N_n(k-1)$ and we will later prove convergence of the terms containing it. We rescale, and split $C_n^{(2)}(k)$ into

$$\begin{aligned} n^{-1/3} C_n^{(2)}(tn^{2/3}) &= n^{-1/3} \sum_{k=1}^{tn^{2/3}} \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [k f_T(0) D_k - f_T(0) \Sigma_{k-1} | \mathcal{E}_{k-1}] \\ &\quad + n^{-1/3} \sum_{k=1}^{tn^{2/3}} \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [f'_T(0) k \Sigma_{k-1} D_k + k o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-1/3}) | \mathcal{E}_{k-1}] + \varepsilon_n, \end{aligned} \quad (5.52)$$

where ε_n represents the terms containing $N_n(k-1)$. Again we rescale and perform a term-by-term analysis.

First term in (5.52). Expanding $(1 - F_T(\Sigma_{k-1}))^{-1}$ gives

$$\frac{f_T(0)}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \mathbb{E} [k S_k - \sum_{j=1}^{k-1} S_j | \mathcal{E}_{k-1}] + \frac{f_T^2(0)}{n^{1/3}} \sum_{k=1}^{tn^{2/3}} O_{\mathbb{P}}(\Sigma_{k-1}) \cdot \mathbb{E} [k D_k - \Sigma_{k-1} | \mathcal{E}_{k-1}], \quad (5.53)$$

The second term is almost surely dominated by the first for n sufficiently large, so that it is enough to show (uniform) convergence of the first term. By Lemma 15,

$$-\frac{1}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} \sum_{j=1}^{k-1} S_j \xrightarrow{\mathbb{P}} -\frac{t^2}{2} \mathbb{E}[S], \quad \frac{1}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} k \mathbb{E}[S_k] \xrightarrow{\mathbb{P}} \frac{t^2}{2} \mathbb{E}[S]. \quad (5.54)$$

Therefore, (5.53) converges to zero in probability. Moreover, the convergence is uniform in $t \leq \bar{t}$ by Lemma 14.

Second term in (5.52). Expanding $(1 - F_T(\Sigma_{k-1}))^{-1}$ and ignoring all but the highest order term, which can be almost surely dominated, we get for the second term

$$f'_T(0)\mathbb{E}[S]n^{-7/3} \sum_{k=1}^{tn^{2/3}} k \sum_{j=1}^{k-1} S_j + n^{-1/3} \sum_{k=1}^{tn^{2/3}} k o_{\mathbb{P}}(n^{-1}) + tn^{1/3} o_{\mathbb{P}}(n^{-1/3}). \quad (5.55)$$

One can check, similarly as in Lemma 15, that $N^{-3} \sum_{k=1}^N k \sum_{j=1}^k S_j$ converges in probability to a non-trivial limit. Since $7/3 > (2/3) \cdot 3$, the first term converges to zero in probability. In addition, it converges uniformly in $t \leq \bar{t}$ because of the monotonicity of the sum. The small-o terms can be uniformly dominated as has already been done in (5.47).

Third term in (5.52). The remaining term is

$$\varepsilon_n = \sum_{k=1}^{tn^{2/3}} \frac{n^{-1/3}}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[N_n(k-1)(f_T(0)D_k + f'_T(0)\Sigma_{k-1}D_k + o_{\mathbb{P}}(n^{-1})) | \mathcal{E}_{k-1}].$$

Again it is sufficient to prove (uniform) convergence for the first term in the Taylor expansion of $(1 - F_T(\Sigma_{k-1}))^{-1}$. This simplifies the previous expression to

$$\frac{f_T(0)\mathbb{E}[S]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} N_n(k-1) + \frac{f'_T(0)\mathbb{E}[S]}{n^{4/3}} \sum_{k=1}^{tn^{2/3}} N_n(k-1)\Sigma_{k-1} + n^{-4/3} \sum_{k=1}^{tn^{2/3}} N_n(k-1)o_{\mathbb{P}}(n^{-1}). \quad (5.56)$$

The second and third terms are again almost surely dominated by the first for n large. Moreover, the first converges to zero uniformly in probability by the following lemma:

Lemma 16. *Let $(N_n(k))_{k \geq 0}$ be the process (3.5) for the General Arrivals Model and $a \in \mathbb{R}^+$. Then,*

$$n^{-2/3} \sup_{j \leq an^{2/3}} |N_n(j)| \xrightarrow{\mathbb{P}} 0. \quad (5.57)$$

Proof. The proof proceeds as in Lemma 9. We split $N_n(j)$ as the sum of a martingale and a predictable process, $N_n(j) = M_n(j) + C_n(j)$, and bound each one separately. The term containing the martingale $M_n(j)$ can be bounded through Doob's inequality, leaving us to bound $\mathbb{E}[M_n^2(an^{2/3})]/(\varepsilon n^{2/3})^2$. As was noticed in Lemma 9, $\mathbb{E}[M_n(k)^2] = \mathbb{E}[V_n(k)]$, where $V_n(k)$ is the predictable quadratic variation of $M_n(k)$, and its expectation is given by

$$\mathbb{E}[V_n(k)] = \mathbb{E}\left[\sum_{i=1}^k (\mathbb{E}[A_n(i)^2 | \mathcal{F}_{i-1}] - \mathbb{E}^2[A_n(i) | \mathcal{F}_{i-1}])\right] \leq \sum_{i=1}^k \mathbb{E}[A_n(i)^2]. \quad (5.58)$$

This term can be easily bounded exploiting Lemma 13. We have

$$\frac{1}{(\varepsilon n^{2/3})^2} \mathbb{E}[V_n(an^{2/3})] \leq \frac{1}{(\varepsilon n^{2/3})^2} \sum_{i=1}^{an^{2/3}} \mathbb{E}[(A'_n(i))^2], \quad (5.59)$$

which tends to zero because $\mathbb{E}[A_n'^2(i)] < \infty$ uniformly in $i = O(n^{2/3})$ by Lemma 13. The $C_n(j)$ term (that was computed in (5.32) and (5.38)) is the difference of two increasing processes, and thus can be easily bounded from above and below as was done in Lemma 8. \square

This concludes the proof that

$$\sup_{t \leq \bar{t}} |n^{-1/3} C_n^{(2)}(tn^{2/3})| \xrightarrow{\mathbb{P}} 0.$$

and thus we have proven that

$$\sup_{t \leq \bar{t}} |n^{-1/3} C_n(n^{2/3}t) - \beta t - t^2 f'_T(0) \mathbb{E}[S]^2/2| \xrightarrow{\mathbb{P}} 0.$$

This completes the proof of (i). \square

5.2.2 Proof of (ii)

First we compute $\mathbb{E}[A_n(k)^2 | \mathcal{F}_{k-1}]$. By proceeding as in (4.43) we obtain

$$\begin{aligned} \mathbb{E}[A_n(k)^2 | \mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) | \mathcal{F}_{k-1}] &= \mathbb{E} \left[\sum_{\substack{l \neq m \\ l, m \notin \nu_k}} \mathbb{1}_{\{\Sigma_{k-1} \leq T_m \leq \Sigma_k\}} \cdot \mathbb{1}_{\{\Sigma_{k-1} \leq T_l \leq \Sigma_k\}} | \mathcal{F}_{k-1} \right] \\ &= \sum_{\substack{l \neq m \\ l, m \notin \nu_k}} \mathbb{E} \left[\frac{F_T(\Sigma_k) - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_{k-1})} \cdot \frac{F_T(\Sigma_k) - F_T(\Sigma_{k-1})}{1 - F_T(\Sigma_{k-1})} | \mathcal{F}_{k-1} \right] \\ &= \frac{\sum_{l \neq m} \mathbb{E}[(f_T(\Sigma_{k-1})D_k + o_{\mathbb{P}}(D_k^{-4/3}))^2 | \mathcal{F}_{k-1}]}{(1 - F_T(\Sigma_{k-1}))^2}, \end{aligned} \quad (5.60)$$

where the sum is over the set $\{l, m : l \neq m, l, m \notin \nu_k\}$ wherever not specified. We also denoted, for convenience, $F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1})D_k$ as $o_{\mathbb{P}}(D_k^{4/3})$. We proceed as in (4.44) and (4.45). By Lemma 10,

$$\begin{aligned} \mathbb{E}[A_n(k)^2 | \mathcal{F}_{k-1}] - \mathbb{E}[A_n(k) | \mathcal{F}_{k-1}] &= \frac{1}{(1 - F_T(\Sigma_{k-1}))^2} \sum_{l \neq m} \mathbb{E}[f_T^2(\Sigma_{k-1})D_k^2 + 2f_T(\Sigma_{k-1})D_k \cdot o_{\mathbb{P}}(D_k^{4/3}) + o_{\mathbb{P}}(D_k^2) | \mathcal{F}_{k-1}] \\ &= \frac{1}{(1 - F_T(\Sigma_{k-1}))^2} \sum_{l \neq m} \left((f_T(0) + f'_T(0)\Sigma_{k-1} + o_{\mathbb{P}}(\Sigma_{k-1}))^2 \mathbb{E}[D_k^2] + o_{\mathbb{P}}(n^{-2}) \right). \end{aligned} \quad (5.61)$$

As usual, here $o_{\mathbb{P}}(\Sigma_{k-1})$ is a shorthand notation for $f_T(\Sigma_{k-1}) - f_T(0) - f'_T(0)\Sigma_{k-1}$. Developing the coefficient of $\mathbb{E}[D_k^2]$ reveals that it has the form $f_T(0)^2 + \alpha_k n^{-1/3} + \beta_k o(n^{-1/3})$, with α_k and β_k converging in probability to a constant, for $k = O(n^{2/3})$. We can ignore all the terms except the one with the leading order, $f_T^2(0)$. From this point onwards the computations are identical to (4.47).

5.2.3 Proof of (iii) and (iv)

The proof of (iii) in the exponentials arrivals case can be carried over to the general arrivals case without any significant changes, since it relies only on (ii) and Lemma 16.

For (iv), we split the quantity according to

$$\begin{aligned} n^{-2/3} \mathbb{E} \left[\sup_{t \leq \bar{t}} |M_n(tn^{2/3}) - M(tn^{2/3}-)|^2 \right] &\leq n^{-2/3} \mathbb{E} \left[\sup_{k \leq \bar{t}n^{2/3}} |A_n(k)|^2 \right] \\ &\quad + n^{-2/3} \mathbb{E} \left[\sup_{k \leq \bar{t}n^{2/3}} |\mathbb{E}[A_n(k) | \mathcal{F}_{k-1}]|^2 \right]. \end{aligned} \quad (5.62)$$

The second term is straightforward. Indeed, (5.32) and (5.37) give the crude bound

$$\mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] \stackrel{\text{a.s.}}{\leq} A_n^{(1)}(k) \stackrel{\text{a.s.}}{\leq} C + o_{\mathbb{P}}(1),$$

for $k = O(n^{2/3})$ and some positive constant $C > 1$, uniform over $k \leq \bar{t}n^{2/3}$. The first term can also be estimated imitating (4.58). Indeed, fix $\varepsilon > 0$ and split it as

$$\begin{aligned} \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} |A_n(k)|^2\right] &= \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \mathbf{1}_{\{\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \leq \varepsilon n^{2/3}\}}\right] \\ &\quad + \mathbb{E}\left[\sup_{k \leq \bar{t}n^{2/3}} A_n(k)^2 \mathbf{1}_{\{\sup_{k \leq \bar{t}n^{2/3}} A(k)^2 > \varepsilon n^{2/3}\}}\right]. \end{aligned} \quad (5.63)$$

The first term is trivially bounded. The bounding of the second term proceeds as in (4.59) and is concluded through Lemma 13. \square

6 Proof of Theorem 8

Note that α in (3.38) is such that $\alpha < 1$. Some simple relations hold between α and ℓ and we will use these throughout this section:

$$\ell - \frac{\alpha}{2} = \alpha\ell, \quad \ell + \frac{\alpha}{2} = \alpha(\ell + 1). \quad (6.1)$$

We need to adapt the model assumptions to the ℓ -th order contact case. Assume that the distribution function of the arrival times satisfies

$$F_T(x) - F_T(\bar{x}) = f_T(\bar{x}) \cdot (x - \bar{x}) + o(|x - \bar{x}|^{1+\alpha/2}). \quad (6.2)$$

This is, for example, the case when $F_T \in \mathcal{C}^2([0, \infty))$. The service times are given by

$$D_i := \frac{S_i}{n}(1 + \beta n^{-\alpha/2}), \quad i \geq 1. \quad (6.3)$$

We assume that the maximum of the density $f_T(\cdot)$ is obtained in zero, and the heavy-traffic condition is then defined as

$$nf_T(0) \cdot \mathbb{E}[D] = 1 + \beta n^{-\alpha/2}. \quad (6.4)$$

6.1 Proof of Theorem 8

We proceed by proving conditions (i)-(iv). We will treat condition (i) in great detail as this changes profoundly, since the limiting drift is significantly different. We will then discuss how (ii)-(iv) follow from the calculations in Section 5.

Proof of (i). The starting point is again (5.32),

$$\begin{aligned} \mathbb{E}[A_n(k)|\mathcal{F}_{k-1}] - 1 &= \frac{n}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[(F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - 1/n)|\mathcal{E}_{k-1}] \\ &\quad - \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E}[(|\nu_k|(F_T(\Sigma_k) - F_T(\Sigma_{k-1})) - F_T(\Sigma_{k-1}))|\mathcal{E}_{k-1}] \\ &=: A_n^{(1)}(k) - A_n^{(2)}(k), \end{aligned} \quad (6.5)$$

and the corresponding drift decomposition

$$C_n(s) = \sum_{k=1}^s (A_n^{(1)}(k) - A_n^{(2)}(k)) =: C_n^{(1)}(s) + C_n^{(2)}(s). \quad (6.6)$$

Recall that $A_n^{(1)}(k)$ contains the terms appearing in the limit, while $A_n^{(2)}(k)$ the ones that vanish. Expanding gives us the equivalent of (5.37), in the form

$$A_n^{(1)}(k) = \frac{1}{1 - F_T(\Sigma_{k-1})} \cdot \left(\beta + \frac{f_T^{(\ell)}(0)}{\ell!} \mathbb{E}[S_1] \frac{(\sum_{j=1}^{k-1} S_j)^\ell}{n^{\ell\alpha}} + o_{\mathbb{P}}(1) \right) \cdot n^{-\alpha/2} \quad (6.7)$$

It is easy to check that both the linear part of the drift and the error term converge uniformly by proceeding as in (5.44) and the following computations. Therefore, we focus on the second term of $A_n^{(1)}(k)$, that is, on

$$\left(\mathbb{E}[S] \frac{f_T^{(\ell)}(0)}{\ell!} \right) n^{-\alpha/2} \sum_{i=1}^{tn^\alpha} \frac{1}{1 - F_T(\Sigma_{i-1})} \left(\frac{\sum_{j=1}^{i-1} S_j}{n} \right)^\ell, \quad (6.8)$$

for which we prove (uniform) convergence in probability. We begin by computing

$$\begin{aligned} \left| \frac{1}{n^{\alpha(\ell+1)}} \sum_{i=1}^{tn^\alpha} \left(\sum_{j=1}^{i-1} S_j \right)^\ell - \frac{t^{\ell+1}}{\ell+1} \mathbb{E}[S]^\ell \right| &= \frac{1}{n^{\alpha(\ell+1)}} \left| \sum_{i=1}^{tn^\alpha} \left(\sum_{j=1}^{i-1} S_j \right)^\ell - \sum_{i=1}^{tn^\alpha} ((i-1)\mathbb{E}[S])^\ell + o(n^{\alpha(\ell+1)}) \right| \\ &\leq \frac{1}{n^{\alpha(\ell+1)}} \sum_{i=1}^{tn^\alpha} \left| \left(\sum_{j=1}^{i-1} S_j \right)^\ell - (i-1)^\ell \mathbb{E}[S]^\ell \right| + o(1) \end{aligned} \quad (6.9)$$

We now make use of the following generalization of Lemma 14:

Lemma 17. *Assume $(S_i)_{i \geq 0}$ is a sequence of i.i.d. random variables with finite second moment. Then for any $\alpha > 0, \beta \in \mathbb{R}$ such that $-\alpha < 2\beta$,*

$$\frac{\sup_{k \leq tn^\alpha} \left| \left(\sum_{i=1}^k S_i \right)^\ell - k^\ell \mathbb{E}[S]^\ell \right|}{n^{\alpha\ell + \beta}} \xrightarrow{\mathbb{P}} 0. \quad (6.10)$$

Proof. The proof is an application of Lemma 14, hence we only sketch it. We have

$$\begin{aligned} \sup_{k \leq tn^\alpha} \left| \left(\sum_{i=1}^k S_i \right)^\ell - k^\ell \mathbb{E}[S]^\ell \right| &\leq \sup_{k \leq tn^\alpha} \left| \left(k\mathbb{E}[S] + \sup_{k \leq tn^\alpha} \left| \sum_{i=1}^k S_i - k\mathbb{E}[S] \right| \right)^\ell - k^\ell \mathbb{E}[S]^\ell \right| \\ &\vee \sup_{k \leq tn^\alpha} \left| \left(k\mathbb{E}[S] - \sup_{k \leq tn^\alpha} \left| \sum_{i=1}^k S_i - k\mathbb{E}[S] \right| \right)^\ell - k^\ell \mathbb{E}[S]^\ell \right|. \end{aligned} \quad (6.11)$$

Both terms on the right can be shown to converge to zero when appropriately rescaled. To do so, it is enough to study the leading order term, which is

$$\sup_{k \leq tn^\alpha} k^{\ell-1} \mathbb{E}[S]^{\ell-1} \sup_{k \leq tn^\alpha} \left| \sum_{i=1}^k S_i - k\mathbb{E}[S] \right| = (tn)^{\alpha(\ell-1)} \mathbb{E}[S]^{\ell-1} \sup_{k \leq tn^\alpha} \left| \sum_{i=1}^k S_i - k\mathbb{E}[S] \right|, \quad (6.12)$$

and this converges to zero when divided by $n^{\alpha\ell + \beta}$, with $\beta > -\alpha/2$. \square

Note that when $\ell = 1$ we do, in fact, recover Lemma 14, since in that case $\alpha + \beta > \alpha - \alpha/2 = \alpha/2$.

Then, by Lemma 17 the right side of (6.9) converges to zero. The convergence is uniform in $t \leq \bar{t}$ by monotonicity in t . We can similarly analyse $A_n^{(2)}(k)$ and $C_n^{(2)}(k)$. Equation (5.51) then becomes

$$C_n^{(2)}(s) = \sum_{k=1}^s \frac{1}{1 - F_T(\Sigma_{k-1})} \times \quad (6.13)$$

$$\times \mathbb{E} \left[|\nu_k| f_T(0) D_k - f_T(0) \Sigma_{k-1} + \frac{f^{(\ell)}(0)}{\ell!} |\nu_k| (\Sigma_{k-1})^\ell D_k + |\nu_k| o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-\alpha/2}) \middle| \mathcal{E}_{k-1} \right],$$

with the usual convention that $o_{\mathbb{P}}(n^{-1}) =: |F_T(\Sigma_k) - F_T(\Sigma_{k-1}) - f_T(\Sigma_{k-1}) D_k|$ and $o_{\mathbb{P}}(n^{-\alpha/2}) =: |F_T(\Sigma_{k-1}) - f_T(0) \Sigma_{k-1}|$. This gives a decomposition of the drift $C_n^{(2)}(k)$ similar to (5.52) as

$$n^{-\alpha/2} C_n^{(2)}(tn^\alpha) = n^{-\alpha/2} \sum_{k=1}^{tn^\alpha} \frac{1}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [k f_T(0) D_k - f_T(0) \Sigma_{k-1} | \mathcal{E}_{k-1}] \quad (6.14)$$

$$+ \sum_{k=1}^{tn^\alpha} \frac{n^{-\alpha/2}}{1 - F_T(\Sigma_{k-1})} \mathbb{E} [f_T^{(\ell)}(0) k (\Sigma_{k-1})^\ell D_k + k o_{\mathbb{P}}(n^{-1}) + o_{\mathbb{P}}(n^{-\alpha/2}) | \mathcal{E}_{k-1}] + \varepsilon_n,$$

where ε_n groups all the terms containing $N_n(k)$. The first term is less straightforward now, because the first order terms in its expansion do not cancel out immediately. Consider (again ignoring the higher order terms in the expansion of $(1 - F_T(\Sigma_{k-1}))^{-1}$)

$$n^{-\alpha/2} \left| \sum_{k=1}^{tn^\alpha} (k f_T(0) \mathbb{E}[D_1] - f_T(0) \Sigma_{k-1}) \right| = f_T(0) n^{-1-\alpha/2} \left| \sum_{k=1}^{tn^\alpha} \left(\sum_{j=1}^{k-1} S_j - k \mathbb{E}[S] \right) \right|$$

$$\leq f_T(0) n^{-1-\alpha/2} \sum_{k=1}^{tn^\alpha} \left| \sum_{j=1}^{k-1} S_j - k \mathbb{E}[S] \right| \quad (6.15)$$

It is now sufficient to apply Lemma 14 with $\beta = 1 - \alpha/2$ (note that $2\beta < \alpha$). Therefore, the first term in (6.14) tends to zero in probability, uniformly in $t \leq \bar{t}$. The remaining terms are treated, without additional complications, in a similar manner as in Section 4.

Proof of (ii). This was proved through an analysis of the leading-order term, in which the ℓ -th derivative of the density played no role.

Proof of (iii). The proof of (iii) relies heavily on Lemma 16 which, in turn, relies on the analysis of the order statistics done in Lemma 13. Since the latter does not depend on the derivatives of the density (but rather on its continuity), the proof carries over.

Proof of (iv). Again, the proof relies on the analysis carried out in Lemma 13.

Having proved conditions (i)-(iv), this concludes the proof of Theorem 8. \square

7 Proof of Theorem 1

The process $Q_n(\cdot)$ approximates the real queueing process $Q_n^\Delta(\cdot)$, because $Q_n(\cdot)$ never idles and is embedded at service completions. Therefore, establishing Theorem 1 from Theorem 5 requires to prove that these differences between $Q_n(\cdot)$ and $Q_n^\Delta(\cdot)$ are asymptotically negligible as $n \rightarrow \infty$. We assume without loss of generality that the exponential random variables $(T_i)_{i=1}^n$ have unit mean (equivalently $\mathbb{E}[S] = 1$).

7.1 Idle times

Recall that, whenever the queue $Q_n(\cdot)$ is empty at the end of a service (say, the k -th service), the customer with the smallest arrival time is drawn from the pool and placed into service. Denote this customer by $c(k)$. Since the minimum of n rate one exponential random variables is again an exponential random variable with rate n , conditioned on $|\nu_k|$, $I_k := T_{c(k)}$ is distributed as

$$I_k \stackrel{d}{=} \frac{\text{Exp}(1)}{n - |\nu_k|}, \quad (7.1)$$

where $\text{Exp}(1)$ is an exponential random variable with rate one. The random variable I_k represents *the time the server would have idled if customer $c(k)$ was not immediately placed into service*. Alternatively, I_k is the idle period after the k -th service in the coupled $\Delta_{(i)}/G/1$ queue. Thus we name I_k a *virtual* idle time. In particular $|\nu_k| = O(n^{2/3})$ for $k = O(n^{2/3})$, and therefore its contribution in (7.1) is negligible and we can think of the virtual idle periods up to time $k = O(n^{2/3})$ as being independent exponential random variables with rate n .

Let $\beta_n(k)$ be the number of customer who have been taken from the population for immediate service before the completion of the $(k+1)$ -th service. Equivalently, $\beta_n(k)$ is the number of idle periods in the coupled $\Delta_{(i)}/G/1$ queue before the completion of the $(k+1)$ -th service. The *cumulative* virtual idle time up to the k -th service completion takes the form

$$\mathcal{I}(k) = \sum_{i=1}^{\beta_n(k)} I_i. \quad (7.2)$$

For exponential arrivals, an explicit expression for the virtual idle periods is available in (7.1), thus we can estimate the average cumulative virtual idle time at step $k = O(n^{2/3})$ to be

$$\sum_{i=1}^{\beta_n(k)} I_i \approx \beta_n(k) \cdot \frac{1}{n} \quad (7.3)$$

We now aim at making (7.3) rigorous by proving that $\mathcal{I}(k)$ is asymptotically negligible, uniformly in $k = O(n^{2/3})$. First we show that $\beta_n(k) = O_{\mathbb{P}}(n^{1/3})$ and to this end we prove the following representation:

Lemma 18. *Under the same assumptions as in Theorem 6,*

$$\beta_n(k) = -\inf_{j \leq k} (N_n(j) \wedge 0), \quad k \geq 0, \text{ a.s.} \quad (7.4)$$

Proof. (7.4) holds for $k = 0$ because in this case both $\beta_n(0) = 0$ and $N_n(0) = 0$. Assume (7.4) holds for $k \geq 1$. Without loss of generality we can also assume that

$$\beta_n(k) = -\inf_{j \leq k} (N_n(j) \wedge 0) = -N_n(k). \quad (7.5)$$

Let \bar{k} be the minimum index such that $\bar{k} > k$, $N_n(\bar{k}-1) = N_n(k)$ and $A_n(\bar{k}) = 0$. Equivalently, at the end of the \bar{k} -th service time there are no customers in the queue (and it is the first time after the k -th service that this happens). By the definition of $\beta_n(\cdot)$, we have that $\beta_n(\bar{k}-1) = \beta_n(k)$ and $\beta_n(\bar{k}) = \beta_n(k) + 1$. On the other hand we have that $N_n(\bar{k}) = N_n(\bar{k}-1) + A_n(\bar{k}) - 1 = N_n(k) - 1$. This gives

$$\beta_n(\bar{k}) = \beta_n(k) + 1 = -N_n(k) + 1 = -N_n(\bar{k}) = -\inf_{j \leq \bar{k}} (N_n(j) \wedge 0). \quad (7.6)$$

Moreover, for every q such that $k < q < \bar{k}$,

$$-\inf_{j \leq q} (N_n(j) \wedge 0) = -N_n(k), \quad (7.7)$$

and $\beta_n(q) = \beta_n(k)$, by definition of \bar{k} . \square

The following lemma shows that $n^{-1/3}\beta_n(tn^{2/3})$ converges in distribution to a non-trivial random variable, hence (7.3) is negligible in the limit when $k = O(n^{2/3})$. Recall that $\beta_n(k)$ denotes the number of customers that have been removed from the population and directly put into service before the end of the $(k+1)$ -st service.

Lemma 19 (Convergence of the number of idle periods). *Fix $t \in (0, \infty)$. Under the same assumptions as in Theorem 6,*

$$n^{-1/3}\beta_n(tn^{2/3}) \xrightarrow{d} -\inf_{s \leq t} (W(s) \wedge 0), \quad (7.8)$$

where $W(\cdot)$ is given in (3.28).

Proof. The operator $\psi : f \mapsto \psi(f)(t) = -\inf_{s \leq t} (f(s) \wedge 0)$ acting from \mathcal{D} to itself is Lipschitz continuous with respect to the Skorohod topology by [35, Theorem 6.1]. Note that

$$n^{-1/3}\beta_n(tn^{2/3}) = \psi(n^{-1/3}N_n(\cdot n^{2/3}))(t). \quad (7.9)$$

Then, since $n^{-1/3}N_n(\cdot n^{2/3}) \xrightarrow{d} W$, by the Continuous Mapping Theorem the image of $n^{-1/3}N_n(\cdot n^{2/3})$ through ψ converge, i.e.,

$$\psi(n^{-1/3}N_n(\cdot n^{2/3})) \xrightarrow{d} \psi(W), \quad (7.10)$$

and this is (7.8). \square

As a direct consequence, the following lemma shows that the cumulative virtual idle time is asymptotically negligible:

Lemma 20 (Convergence of the cumulative idle time). *Assume that the arrival clocks $(T_i)_{i=1}^n$ are exponentially distributed. Fix $t \in (0, \infty)$. Then, under the same assumptions as in Theorem 5 and conditioned on $\{Q_n(s), s \in (0, tn^{2/3})\}$,*

$$\frac{n}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} I_i \xrightarrow{\mathbb{P}} 1. \quad (7.11)$$

Proof. As was noted in (7.1), I_i is distributed as an exponential random variable with rate $n - |\nu_{k_i}|$, where k_i is the time step corresponding to the i -th customer being placed directly into service. We rewrite the sum as

$$\begin{aligned} \frac{n}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} \frac{E_i}{n - |\nu_{k_i}|} &= \frac{1}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} \frac{E_i}{1 - \frac{|\nu_{k_i}|}{n}} \\ &= \frac{1}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} E_i + \frac{1}{\beta_n(tn^{2/3})} \sum_{i=1}^{\beta_n(tn^{2/3})} E_i \frac{|\nu_{k_i}|}{n} + \varepsilon_n, \end{aligned} \quad (7.12)$$

where $\varepsilon_n = o_{\mathbb{P}}(1)$ and $(E_i)_{i \geq 1}$ are i.i.d. exponential random variables with rate 1. By Lemma 19, $\beta_n(tn^{2/3}) \geq Mn^\alpha$ w.h.p. for a fixed $M > 0$ and $\alpha \in (0, 1/3)$. By the LLN, the first term in (7.12) converges in probability to 1, and by Lemma 9 the second term (and consequently the error term) converges to zero. \square

Lemma 20 intuitively says that the total virtual idle time up to time $tn^{2/3}$ is of the same order of magnitude of the number of virtual idle periods up to time $tn^{2/3}$ times the average interarrival time. In particular, as we prove below, $\mathcal{I}(tn^{2/3}) = o_{\mathbb{P}}(1)$, that is, the cumulative virtual idle time up to times of the order $n^{2/3}$ is negligible:

Corollary 2 (Cumulative idle time is negligible). *Fix $T > 0$. Then,*

$$\sup_{t \leq T} \mathcal{I}(tn^{2/3}) \xrightarrow{\mathbb{P}} 0. \quad (7.13)$$

Proof. By monotonicity, $\sup_{t \leq T} \mathcal{I}(tn^{2/3}) = \mathcal{I}(T)$. Fix $\varepsilon > 0$. Then, by Lemma 20

$$\mathcal{I}(Tn^{2/3}) \leq (1 + \varepsilon)\beta_n(tn^{2/3})/n \quad (7.14)$$

with high probability. By Lemma 19, $\beta_n(tn^{2/3})/n \xrightarrow{\mathbb{P}} 0$, so that

$$\mathcal{I}(Tn^{2/3}) \xrightarrow{\mathbb{P}} 0, \quad (7.15)$$

as desired. \square

7.2 Proof of Theorem 1

We start by defining the time change $\varphi_n(\cdot)$. Fix a realization of the arrival and service times $(S_i)_{i=1}^n = (S_i(\omega))_{i=1}^n$ and $(T_i)_{i=1}^n = (T_i(\omega))_{i=1}^n$, where $\omega \in \Omega$, the sample space. We assume throughout that $\mathbb{E}[S] = 1$. We define $\varphi_n(t)$ piece-wise, depending whether a customer is in service at time t or the server is idling at time t .

Because the time scaling in $Q_n^\Delta(\cdot n^{-1/3})$ is $n^{-1/3}$, for the two processes $Q_n^\Delta(\cdot n^{-1/3})$ and $Q_n(\varphi_n(\cdot)n^{2/3})$ to be on a comparable time scale, the time change $\varphi_n(\cdot)$ should be such that

$$\varphi_n(\cdot) = n^{-2/3}\phi_n(\cdot n^{-1/3}), \quad (7.16)$$

for a suitable $\phi_n(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$. We now provide a precise expression of $\phi_n(\cdot)$. We remark that many other choices for $\phi_n(\cdot)$ would work for proving Lemma 3 and Lemma 4, and the one we present here is only the simplest one.

To increase readability, define

$$\vartheta_{k_2}(k_1) := \frac{\sum_{i=1}^{k_1} S_i}{n} + \mathcal{I}(k_2). \quad (7.17)$$

First assume that at time t a service (say, the k -th service) is taking place, then

$$\phi_n(t) = (k-1) + \frac{1}{S_k/n}(t - \nu_k(k-1)) \quad \text{for } t \in [\vartheta_k(k-1), \vartheta_k(k)]. \quad (7.18)$$

Note that, since the queue is serving at time t , $\mathcal{I}(k-1) = \mathcal{I}(k)$. In other words $\phi_n(t)$ is the line joining the points $(\vartheta_k(k-1), k-1)$ and $(\vartheta_k(k), k)$, where the term $\mathcal{I}(k)$ in $\vartheta_k(k)$ takes into account previous idle periods which might have occurred before t .

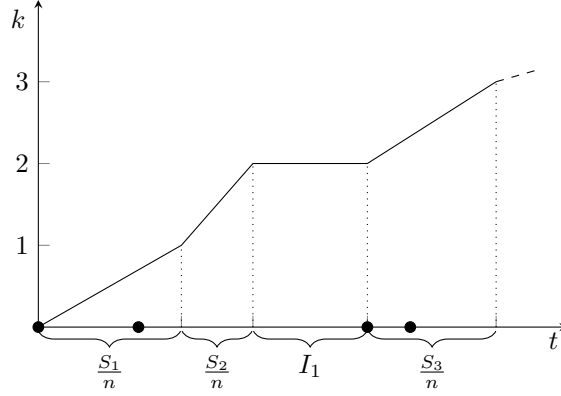


Figure 3: An example of the time change $\phi_n(\cdot)$. The \bullet 's indicate arrival times of customers.

Assume now that an idle period (say, the $(\beta(k) + 1)$ -th idle period) is under way at time t . Then $\phi_n(t)$ takes the form

$$\phi_n(t) = k \quad \text{for } t \in [\vartheta_k(k), \vartheta_k(k) + I_{\beta(k)+1}]. \quad (7.19)$$

In other words, $\phi_n(t)$ is constant during an idle period. This is because $\phi_n(t)$ (being the input of $Q_n(\cdot)$) represents the *number of completed services*. Again, here $\sum_{i=1}^k S_i/n$ takes previous services that have occurred before t into account. Summarizing, $\phi_n(\cdot)$ takes the form

$$\phi_n(t) = \begin{cases} \left(k - 1 + \frac{n}{S_k}(t - \vartheta_k(k-1)) \right) & \text{for } t \in [\vartheta_k(k-1), \vartheta_k(k)] \text{ for some } k; \\ k & \text{for } t \in [\vartheta_k(k), \vartheta_k(k) + I_{\beta(k)+1}] \text{ for some } k. \end{cases} \quad (7.20)$$

In particular $\phi_n(\cdot)$ (and therefore also $\varphi_n(\cdot)$) is a piecewise linear continuous function. See Figure 3 for one possible sample path of $\phi_n(\cdot)$. We now focus on $\varphi_n(\cdot)$ in (7.16) and show that it converges to the identity function. The time change $\varphi_n(t)$ takes the form

$$\varphi_n(t) = \begin{cases} n^{-2/3} \left(k - 1 + \frac{n}{S_k} \left(\frac{t}{n^{1/3}} - \vartheta_k(k-1) \right) \right) & \text{for } \left(\frac{t}{n^{1/3}} \right) \in [\vartheta_k(k-1), \vartheta_k(k)]; \\ n^{-2/3} k & \text{for } \left(\frac{t}{n^{1/3}} \right) \in [\vartheta_k(k), \vartheta_k(k) + I_{\beta(k)+1}]. \end{cases} \quad (7.21)$$

Note that the only values of k for which $\phi_n(t)$ has a meaningful limit as $n \rightarrow \infty$ are $k = O(n^{2/3})$. This observation is consistent with the fact that the original scaling for $Q_n(\cdot)$ is $Q_n(\cdot n^{2/3})$, that is, in order to obtain a meaningful limit we observe the queue length at times when $O(n^{2/3})$ services have been completed. As a consequence, we assume that $k = sn^{2/3}$ for some $s \in \mathbb{R}^+$. Summarizing, the final form of $\varphi_n(t)$ is

$$\varphi_n(t) = \begin{cases} s - \frac{1}{n^{2/3}} + \frac{1}{S_{sn^{2/3}}}(t - \vartheta_{sn^{2/3}}(sn^{2/3} - 1)) & \text{for } t \in n^{1/3}[\vartheta_{sn^{2/3}}(sn^{2/3} - 1), \vartheta_{sn^{2/3}}(sn^{2/3})]; \\ s & \text{for } t \in n^{1/3}[\vartheta_{sn^{2/3}}(sn^{2/3}), \vartheta_{sn^{2/3}}(sn^{2/3}) + I_{\beta(sn^{2/3})}]. \end{cases} \quad (7.22)$$

We now turn to proving the first claim, as formulated in Lemma 3.

Proof of Lemma 3. First note that

$$\sup_{t \leq T} \left| -\frac{1}{n^{2/3}} + \frac{1}{S_{sn^{2/3}}}(t - \vartheta_{sn^{2/3}}(sn^{2/3} - 1)) \right| \leq \frac{2}{n^{2/3}} \rightarrow 0, \quad n \rightarrow \infty, \quad (7.23)$$

implying that we can treat $\varphi_n(t)$ as piece-wise constant, $\varphi_n(t) \equiv s$ on intervals of the form $[l(s), u(s)]$. We now prove (3.16). Since the function $t \mapsto t - \varphi_n(t) = t - s$ (for some fixed s) defined on an interval $[l, u] = [l(s), u(s)]$ is linear in t , it obtains its maximum either in l or u . This implies

$$\lim_{n \rightarrow \infty} \|t - \varphi_n(t)\|_T \leq \lim_{n \rightarrow \infty} \left(\sup_{s \leq S} \left| \vartheta_{sn^{2/3}}(sn^{2/3} - 1) - s \right| \vee \sup_{s \leq S} \left| \vartheta_{sn^{2/3}}(sn^{2/3}) + I_{\beta(sn^{2/3})} - s \right| \vee \sup_{s \leq S} \left| \vartheta_{sn^{2/3}}(sn^{2/3}) - s \right| \right), \quad (7.24)$$

where $x \vee y := \max\{x, y\}$ and $S = S(T) = \frac{T}{\mathbb{E}[S]}(1 + \varepsilon) > 0$ for some $\varepsilon > 0$. The inequality comes from the fact that the three supremums are taken over a larger set, by the definition of $S(T)$. We prove convergence to zero of one of the three terms on the right in (7.24), the others being analogous. Applying the triangle inequality to the last term yields

$$\sup_{s \leq S} \left| \frac{\sum_{i=1}^{sn^{2/3}} S_i}{n^{2/3}} + n^{1/3} \mathcal{I}(sn^{2/3}) - s \right| \leq \sup_{s \leq S} \left| \frac{\sum_{i=1}^{sn^{2/3}} S_i}{n^{2/3}} - s \right| + \sup_{s \leq S} \left| n^{1/3} \mathcal{I}(sn^{2/3}) \right|. \quad (7.25)$$

The first term converges to zero in probability by the functional Law of Large Numbers. The second term converges to zero in probability by Corollary 2. Indeed, the proof of Lemma 20 and Corollary 2 show that $\mathcal{I}(sn^{2/3}) = O_{\mathbb{P}}(n^{-2/3})$. This concludes the proof of (3.16).

We now turn to proving (3.17). Note that $\varphi_n(\cdot) \xrightarrow{\mathbb{P}} \text{id}$, the identity map on $[0, T]$. In particular, id is deterministic, so that by [36, Theorem 11.4.5],

$$(n^{-1/3} Q_n(\cdot n^{2/3}), \varphi_n) \xrightarrow{d} (W, \text{id}). \quad (7.26)$$

This allows to prove claim (i), as follows. By Skorokhod's representation theorem and (7.26) there exist $\bar{Q}_n(\cdot)$, $\bar{\varphi}_n(\cdot)$ and $\bar{W}(\cdot)$ defined on the same probability space $\bar{\Omega}$ such that

$$(\bar{Q}_n(\cdot n^{2/3}), \bar{\varphi}_n) \stackrel{d}{=} (Q_n(\cdot n^{2/3}), \varphi_n) \quad (7.27)$$

and

$$(n^{-1/3} \bar{Q}_n(\cdot n^{2/3}), \bar{\varphi}_n) \xrightarrow{\text{a.s.}} (\bar{W}(\cdot), \text{id}). \quad (7.28)$$

We now dominate (3.17) using the random variables provided by the representation theorem, as follows:

$$\begin{aligned} n^{-1/3} \|\bar{Q}_n(\cdot n^{2/3}) - \bar{Q}_n(\varphi_n(\cdot) n^{2/3})\|_T \\ \leq n^{-1/3} (\|\bar{Q}_n(\cdot n^{2/3}) - \bar{W}(\cdot)\|_T + \|\bar{W}(\cdot) - \bar{Q}_n(\varphi_n(\cdot) n^{2/3})\|_T). \end{aligned} \quad (7.29)$$

Since the limiting process $\bar{W}(\cdot)$ is almost surely continuous, by standard arguments both the convergence $n^{-1/3} \bar{Q}_n(\cdot n^{2/3}) \xrightarrow{\text{a.s.}} \bar{W}(\cdot)$ and $n^{-1/3} \bar{Q}_n(\varphi_n(\cdot) n^{2/3}) \xrightarrow{\text{a.s.}} \bar{W}(\cdot)$ hold with respect to the uniform topology. In particular, both terms on the right in (7.29) converge to zero in probability. Moreover, since $(\bar{Q}_n(\cdot n^{2/3}), \bar{\varphi}_n(\cdot)) \stackrel{d}{=} (Q_n(\cdot n^{2/3}), \varphi_n(\cdot))$,

$$n^{-1/3} \|\bar{Q}_n(\cdot n^{2/3}) - \bar{Q}_n(\varphi_n(\cdot) n^{2/3})\|_T \stackrel{d}{=} n^{-1/3} \|Q_n(\cdot n^{2/3}) - Q_n(\varphi_n(\cdot) n^{2/3})\|_T, \quad (7.30)$$

so that

$$n^{-1/3} \|Q_n(\cdot n^{2/3}) - Q_n(\varphi_n(\cdot) n^{2/3})\|_T \xrightarrow{\mathbb{P}} 0, \quad (7.31)$$

as desired. This concludes the proof of claim (i) and thus of Lemma 3. \square

The fact that during one service the number of arrivals is asymptotically small is crucial in proving that $Q_n(\cdot)$ and $Q_n^\Delta(\cdot)$ are close in the supremum norm. This is the subject of Lemma 4.

Proof of Lemma 4. Note that if $t/n^{1/3} = \sum_{i=1}^k S_i/n + \mathcal{I}(k)$ for $k \in \mathbb{N}$ (i.e. if a service is completed in t , or an idle period is undergoing in t), then

$$Q_n(\phi_n(t)) = Q_n^\Delta(t). \quad (7.32)$$

This follows from the definition of the time change $\phi_n(t)$, as discussed above. This implies that during any service time $Q_n(\phi_n(t))$ and $Q_n^\Delta(t)$ differ by the number of arrivals that have occurred *during that* service time. Moreover, during any idle time $Q_n(\phi_n(t))$ and $Q_n^\Delta(t)$ are both equal to zero. These considerations imply

$$\sup_{t \leq T} |Q_n(\phi_n(\cdot)n^{2/3}) - Q_n^\Delta(\cdot n^{-1/3})| = \max_{k \leq Tn^{2/3}} A_n(k). \quad (7.33)$$

Hence (3.19) implies (3.20). Let now $\varepsilon > 0$ be arbitrary. Then

$$\mathbb{P}(n^{-1/3} \max_{k \leq Tn^{2/3}} A_n(k) \geq \varepsilon) = \mathbb{P}(n^{-1/3} \max_{k \leq Tn^{2/3}} A_n(k) \mathbb{1}_{\{A_n(k) > \varepsilon n^{1/3}\}} \geq \varepsilon). \quad (7.34)$$

In other words, only the very large values of A_n contribute to the probability being computed. By Chebyshev's inequality,

$$\begin{aligned} \mathbb{P}(n^{-1/3} \max_{k \leq Tn^{2/3}} A_n(k) \geq \varepsilon) &\leq \frac{\mathbb{E}[\max_{k \leq Tn^{2/3}} A_n^2(k) \mathbb{1}_{\{A_n(k) > \varepsilon n^{1/3}\}}]}{(n^{1/3}\varepsilon)^2} \\ &\leq \sum_{k=1}^{Tn^{2/3}} \frac{\mathbb{E}[A_n^2(k) \mathbb{1}_{\{A_n(k) > \varepsilon n^{1/3}\}}]}{n^{2/3}\varepsilon^2}. \end{aligned} \quad (7.35)$$

The almost sure domination $A_n(k) \stackrel{\text{a.s.}}{\leq} A'_n(k) := \sum_{i=1}^n \mathbb{1}_{\{T_i \leq S_k/n\}}$ (note that this domination holds for all $k \leq Tn^{2/3}$ simultaneously) gives

$$\begin{aligned} \mathbb{P}(n^{-1/3} \max_{t \leq T} A_n(tn^{2/3}) \geq \varepsilon) &\leq \sum_{k=1}^{Tn^{2/3}} \frac{\mathbb{E}[A_n'^2(k) \mathbb{1}_{\{A'_n(k) > \varepsilon n^{1/3}\}}]}{n^{2/3}\varepsilon^2} \\ &= \frac{T}{\varepsilon^2} \mathbb{E}[A_n'^2(1) \mathbb{1}_{\{A'_n(1) > \varepsilon n^{1/3}\}}]. \end{aligned} \quad (7.36)$$

The right-most term in (7.36) tends to zero because $A_n'^2$ is stochastically dominated by a uniformly integrable random variable, as was proven in Lemma 8. \square

8 Extended discussion

8.1 On other regimes of criticality

Throughout this paper, we have focused on the critical regime, when $\nu := \sup_{t \geq 0} f_T(t)$ is such that $\nu = \mu$, where $\mu = 1/\mathbb{E}[S]$. This is by far the least studied case in the literature, but other cases may be of interest too. In this section, we discuss some known results also in relation to our own results. We denote by t_{\max} the argmax of $f_T(t)$.

The case $\nu > \mu$ has been studied extensively in [17]. When $\nu > \mu$, the queue length grows linearly in a neighborhood of t_{\max} as has been discussed in Section 1.1. In particular, it is possible to prove that there exists a constant $m > 0$ such that the maximum of the fluid-scaled queue length process converges to m , meaning that the queue during ‘peak hour’ grows linearly.

When $\nu = \mu$, t_{\max} can be either zero, or greater than zero. If it is zero, and if $f'_T(0) < 0$, then our results apply, and thus the correct scaling is $n^{-1/3}$ for the space, $n^{-1/3}$ for the time around t_{\max} . If the first derivative (or possibly all the derivatives up to a finite order) of f_T is zero in zero, by appropriately modifying the scalings our approach still yields the correct limit, as was shown in Section 3.2.1. Here ‘correct’ means that both a deterministic drift indicating a depletion-of-points effect and a Brownian contribution appear. If t_{\max} is greater than zero, and $f_T(t)$ is sufficiently smooth, then we must have that $f'_T(t_{\max}) = 0$, suggesting that one has to investigate higher order derivatives as was done in Section 3.2.1 (note that the first non-zero derivative will necessarily be even, provided $f'_T(t)$ is smooth). If $f_T(t)$ is not smooth, i.e., $f'_T(t)$ is discontinuous in t_{\max} , an even wider range of behaviors will possibly be displayed, due to the fact that $f_T(t) - \mu$ might have different contact orders in t_{\max}^- and t_{\max}^+ .

The uniform arrivals case (uniform in $[0, 1]$ say) is special, since all the derivatives of $f_T(t)$ are identically zero. In some sense, it interpolates between our result and [17]. The queue is critical during its entire lifetime, and thus it never grows linearly, similarly to our model. On the other hand, the uniform arrivals model does not lead to a downward drift. Intuitively, this is because there is no ‘tipping point’ after which the intensity of the arrival process diminishes and therefore the depletion-of-points effect does not occur. By observing the queue at a time scale of order $O(1)$, it displays diffusion-type fluctuations (of the order $n^{1/2}$), which end abruptly in 1, when customers stop joining and thus the queue behaves deterministically, with the server flushing out the customers still in the queue. The correct (optimal) scaling in this case has been first identified in [20] (Example 3, p. 365), and can be treated using the general theory developed in [17]. Indeed, [17, Theorem 1] implies that $Q_n(t)/n \xrightarrow{d} 0$, while [17, Theorem 2] implies that $Q_n(t)/\sqrt{n}$ converges to $X^*(t)$, the reflected version of

$$X(t) = \begin{cases} B^0(t) - \sigma B(t), & t \in [0, 1], \\ 0, & t > 1, \end{cases}$$

where $B^0(t)$ is a standard Brownian Bridge.

The case $\nu < \mu$ is, from the scaling perspective, the least interesting one. The service rate is so large, compared to the instantaneous arrival rate $f_T(t)$, that the queue is empty most of the time and in no instant it builds up as a power of n . One is then tempted to conjecture that, for fixed t , the queue length $Q_n(t)$ converges in distribution to a proper random variable $Q(t)$. Moreover, for $s < t$, $Q_n(\cdot)$ will be zero w.h.p. at some time $\zeta \in (s, t)$, suggesting that $Q(s)$ is close to independent from $Q(t)$. Since locally the arrival process is roughly a Poisson process with rate $f_T(t)$, it is reasonable to also conjecture that $Q(t)$ is the stationary distribution of an $M/G/1$ queue with arrival rate $f_T(t)$. In fact, this has already been proven for the $M_t/M_t/1$ queue in [27] and, using analogous arguments, it is also possible to prove it for the $\Delta_{(i)}/G/1$ model with exponential arrivals.

We conclude by another connection between the present work and [17]. In [17], the queue length process Q_n is such that $Q_n = \phi(X_n)$, where

$$X_n(t) = (A_n(t) - nF_T(t)) - (S^n(B_n(t)) - \mu nB^n(t)) + n(F_T(t) - \mu t). \quad (8.1)$$

Here $A_n(t) := \sum_{i=1}^n \mathbb{1}_{\{T_i \leq t\}}$, F_T is the distribution function of T , S^n is the renewal process associated with the (rescaled) service times $S_1/n, S_2/n, \dots$ and $B_n := \int_0^t \mathbb{1}_{\{Q_n(s) > 0\}} ds$ is the busy time process. Then, for example, [17, Theorem 1] is proved by applying the *continuous-mapping* approach to the fluid-scaled process X_n/n . Let us now scale the process X_n as

$$\begin{aligned} n^{-1/3} X^n(tn^{-1/3}) &= n^{2/3} \left(\frac{A_n(tn^{-1/3})}{n} - F_T(tn^{-1/3}) \right) - n^{2/3} \left(\frac{S^n(B_n(tn^{-1/3}))}{n} - \mu B_n(tn^{-1/3}) \right) \\ &\quad + n^{2/3} (F_T(tn^{-1/3}) - \mu tn^{-1/3}). \end{aligned} \quad (8.2)$$

It is immediate that, if $F_T(x) \approx f_T(0)x + \frac{f_T'(0)}{2}x^2$, the third term in (8.2) yields the quadratic drift. Moreover, it seems plausible that the first and second term converge to (independent) Brownian motions, respectively through a martingale FCLT (or generalized Donsker's Theorem) and through a FCLT for renewal processes. Since the sum of two independent Brownian motions is a Brownian motion itself, the limit of $n^{-1/3} X^n(tn^{-1/3})$ coincides with our result.

8.2 On the relationship between the queue and random graphs

The queueing system in this paper can be used to generate a directed random graph by saying that customer i is connected to customer j when i arrives during the service period of j . Assume the customers have an exponential clock. Conditionally on customer i not having joined the queue yet and conditionally on the service time D_j of the presently served customer j , customer i will be connected to customer j with probability

$$p_{ij} = 1 - e^{-D_j} = 1 - e^{-S_j/n}. \quad (8.3)$$

In particular, when the service times are deterministic, i.e., $S_j \equiv \lambda$, these probabilities are precisely equal to the edge probabilities of the Erdős-Rényi random graph (see [6, 15, 22] for monographs on random graphs) with parameter $1 - e^{-\lambda/n} \approx \lambda/n$. In this case, Aldous [2] has shown that the scaling limit in Theorem 6 appears precisely when $\lambda = 1$, and with the constant $f_T'(0)/f_T(0)^2$ replaced by -1 . This follows, intuitively, by noting that the depletion-of-points effects in the two problems is the same. Note however that the process described by (3.3) is not the graph construction itself, but is instead equivalent to the exploration of the connected components of the graph. Indeed, edges in the graph that form cycles are not explored (i.e. a customer in the queue is not allowed to rejoin the queue), and thus also not observed, so that we obtain a tree rather than a graph.

When S_j is not a.s. constant, then the edge probabilities in (8.3) are not symmetric. We can think of this as the cluster exploration of a *directed* random graph, where, due to the depletion-of-points effect, at most one of the two possible edges between two vertices will ever be found (i.e. for each pair of customers i and j , either i joins the queue while j is being served, or vice versa, or neither of the two). The service time random variables $(S_j)_{j=1}^n$ create *inhomogeneity* in the graph, of a form that has not been investigated in the random graph community.

If, instead of our assumptions, we assume that the clock of customer i has a rate proportional to D_i (i.e., when the job size is larger, then it will join the queue more quickly, inversely proportional to the job size), then the edge probabilities become

$$p_{ij} = 1 - e^{-\lambda D_i D_j} = 1 - e^{-\lambda S_i S_j / n^2}. \quad (8.4)$$

Taking $\lambda = an$, this becomes $p_{ij} = 1 - e^{-aS_i S_j / n}$. This is closely related to the so-called *generalized random graph* with i.i.d. vertex weights, as studied in [3, 4, 14, 23]. See also [7] for

the most general setting, and [8, 9, 33] for related models. Interestingly, a *size-biasing* takes place in the form that the vertices that are found by the exploration process have weights with the size-biased distribution S_j^* given by

$$\mathbb{P}(S_j^* \leq x) = \frac{\mathbb{E}[S_j \mathbb{1}_{\{S_j \leq x\}}]}{\mathbb{E}[S]}. \quad (8.5)$$

In the queueing setting, this means that, if $c(i)$ is the i -th customer joining the queue, the service time of $c(i)$, $S_{c(i)}$ is not distributed as S . In [7] it is shown that the cluster exploration obeys the scaling in Theorem 6 precisely when $\mathbb{E}[S^3] < \infty$, while different scalings can occur when $\mathbb{E}[S^3] = \infty$ and S has a power-law distribution. Since the size-biasing is absent in our work, our assumption $\mathbb{E}[S^2] < \infty$ corresponds to the assumption $\mathbb{E}[S^3] < \infty$ in [7]. This suggests that similar scaling limits as the ones in [3, 23] (in which the exploration process for inhomogeneous random graphs with $\mathbb{E}[S^3] = \infty$ was shown to converge) might appear for queues with vanishing populations of customers when S has a power-law distribution with $\mathbb{E}[S^2] = \infty$. This is an interesting problem for future research.

8.3 On the relationship between the queue and the Grenander estimator

The arrival process in (4.11) can be seen as (n times) the empirical distribution function for n trials of the random variable T . This seemingly trivial observation reveals a connection between our result Theorem 6 and some results in the statistics literature. Suppose we are given a random variable T whose law is supported on $[0, 1]$ and admits a positive density $f_T(t)$ such that $\inf_{t \in [0, 1]} |f_T(t)| > 0$ and $t \mapsto f_T(t)$ is non-increasing. Given that the distribution of T is sampled n times as T_1, \dots, T_n , our task is to estimate f_T . In [12] the so-called Grenander estimator is introduced as an estimator of f_T . It is constructed as follows. Let $F_T^{(n)}(t) := (1/n) \sum_{i=1}^n \mathbb{1}_{\{T_i \leq t\}}$ be the empirical distribution function of the given sample and let $\hat{F}_T^{(n)}(t)$ be the concave conjugate of $F_T^{(n)}(t)$. Recall that the concave conjugate is the (pointwise) smallest concave function such that

$$\hat{F}_T^{(n)}(t) \geq F_T^{(n)}(t), \quad \forall t \in [0, 1]. \quad (8.6)$$

Then, the Grenander estimator $\hat{f}_T^{(n)}$ is defined as the left derivative of $\hat{F}_T^{(n)}(t)$. It turns out that studying the *inverse process* $U_n(a) = \sup\{t \in [0, 1] : F_T^{(n)}(t) - at \text{ is maximal}\}$ is equivalent to studying $\hat{f}_T^{(n)}(t)$, but considerably easier. In particular, in [13] the task of estimating the (rescaled) L^1 error of the estimator, $n^{1/3} \|\hat{f}_T^{(n)} - f_T\|_{L^1([0, 1])}$, is proven to be equivalent to studying $\int |V_n^E(a)| da$, with $V_n^E(a) := n^{1/3}(U_n(a) - g(a))$ and $g(a)$ the inverse of f_T . It is readily seen that V_n^E can be rewritten as

$$V_n^E(a) = \sup\{t \in \mathcal{T} : D_n^E(a, t) - n^{1/3}at \text{ is maximal}\}, \quad (8.7)$$

for some appropriate set $\mathcal{T} = \mathcal{T}_n(a)$. Here D_n^E is the process

$$D_n^E(a, t) := n^{1/6}(E_n(g(a) + n^{-1/3}t) - E_n(g(a))) + n^{2/3}(F_T(g(a) + n^{-1/3}t) - F_T(g(a))), \quad (8.8)$$

where $E_n(x) = \sqrt{n}(F_T^{(n)}(x) - F_T(x))$. The problem in [13] of computing the error of the Grenander estimator then reduces to studying a certain complicated functional of the process

$$t \mapsto D_n^E(a, t) - n^{1/3}at. \quad (8.9)$$

Notice that this can be interpreted as the queue length process associated with a system in which the service rate is deterministic, $1/S \equiv a$ and the arrival process is given by $t \mapsto F_T^{(n)}(g(a) + t) - F_T^{(n)}(g(a))$. Moreover, note that the derivative of F_T in $g(a)$ is

$$F_T'(g(a)) = f_T(g(a)) = a = 1/S, \quad (8.10)$$

since g is the inverse of f_T . Remarkably, (8.10) as seen from our perspective is nothing but the criticality assumption in the point $g(a)$. In [13, Theorem 3.2] the authors prove a convergence result for V_n^E of the type of Theorem 6, the most important difference being that the functional of interest in our case is the reflection mapping, while they are interested in the (supremum among all of the) argmax. Their technique is similar to ours, in the sense that in order to prove convergence of the functional they first prove convergence of the underlying process. This is a two-sided Brownian motion in \mathbb{R} , $(B(t))_{t \in \mathbb{R}}$ originating from zero, i.e. $B(0) = 0$. Whether there is an equivalent, in this setting, of our ℓ -th order contact is an interesting question, and one which remains open for future research.

Acknowledgments

This work is supported by the NWO Gravitation NETWORKS grant 024.002.003. The work of RvdH is further supported by the NWO VICI grant 639.033.806. The work of JvL is further supported by an NWO TOP-GO grant and by an ERC Starting Grant.

References

- [1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Corporation, 1964.
- [2] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, **25**(2):812–854, 1997.
- [3] S. Bhamidi, R. van der Hofstad, and J.S.H. van Leeuwaarden. Novel scaling limits for critical inhomogeneous random graphs. *Ann. Probab.*, **40**(6):2299–2361, 2012.
- [4] S. Bhamidi, R. van der Hofstad, and J.S.H. van Leeuwaarden. Scaling limits for critical inhomogeneous random graphs with finite third moments. *Electron. J. Probab.*, 15:no. 54, 1682–1702, 2010.
- [5] P. Billingsley. *Convergence of Probability Measures*, volume **493**. John Wiley & Sons, 2009.
- [6] B. Bollobás. *Random Graphs*, volume **73** of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.
- [7] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, **31**(1):3–122, 2007.
- [8] T. Britton, M. Deijfen, and A. Martin-Löf. Generating simple random graphs with prescribed degree distribution. *J. Stat. Phys.*, **124**(6):1377–1397, 2006.

- [9] F. Chung and L. Lu. *Complex Graphs and Networks*, volume **107** of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2006.
- [10] Herbert A. D. and Haikady N. N. Order statistics. *NJ: John Wiley & Sons*, 7, 2003.
- [11] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1985.
- [12] U. Grenander. On the theory of mortality measurement: Part II. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956.
- [13] P. Groeneboom, G. Hooghiemstra, H. P. Lopuhaä, et al. Asymptotic normality of the L^1 error of the grenander estimator. *The Annals of Statistics*, 27(4):1316–1347, 1999.
- [14] R. van der Hofstad. Critical behavior in inhomogeneous random graphs. *Random Structures and Algorithms*, **42**(4):480–508, 2013.
- [15] R. van der Hofstad. *Random Graphs and Complex Networks, Vol I*. In preparation 2014.
- [16] R. van der Hofstad, A.J.E.M. Janssen, and J. S.H. van Leeuwen. Critical epidemics, random graphs, and brownian motion with a parabolic drift. *Advances in Applied Probability*, 42(4):1187–1206, 2010.
- [17] H. Honnappa, R. Jain, and A. R. Ward. A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems*, 80(1-2):71–103, 2014.
- [18] Harsha Honnappa, Rahul Jain, and Amy R Ward. On transitory queueing. *arXiv preprint arXiv:1412.2321*, 2014.
- [19] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, 2(1):150–177, 1970.
- [20] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. II. *Advances in Applied Probability*, 2(1):355–369, 1970.
- [21] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes; 2nd ed.* Grundlehren der mathematischen Wissenschaften A Series of Comprehensive Studies in Mathematics. Springer, Berlin, 2003.
- [22] S. Janson, T. Łuczak, and A. Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.
- [23] A. Joseph. The component sizes of a critical random graph with pre-described degree sequence. *The Annals of Applied Probability*, 24(6):2560–2594, 2014.
- [24] J. B. Keller. Time-dependent queues. *SIAM Review*, 24(4):401–412, 1982.
- [25] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, 2007.
- [26] G. Louchard. Large finite population queueing systems. the single-server model. *Stochastic Processes and their Applications*, 53(1):117–145, 1994.

- [27] A. Mandelbaum and W. A. Massey. Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20(1):33–64, 1995.
- [28] A. Martin-Löf. The final size of a nearly critical epidemic, and the first passage time of a wiener process to a parabolic barrier. *Journal of applied probability*, 35(3):671–682, 1998.
- [29] W. A. Massey. Non-stationary queues. *PhD thesis*, 1982.
- [30] W. A. Massey. Asymptotic analysis of the time dependent M/M/1 queue. *Mathematics of Operations Research*, 10(2):pp. 305–327, 1985.
- [31] G. F. Newell. Queues with time-dependent arrival rates I, II and III. *Journal of Applied Probability*, 5(3):436–451 (I); 436–451 (II); 591–606 (III), 1968.
- [32] G. F. Newell. Applications of queueing theory. Technical report, 1982.
- [33] I. Norros and H. Reittu. On a conditionally Poissonian graph process. *Adv. in Appl. Probab.*, **38**(1):59–75, 2006.
- [34] A. V. Skorokhod. Limit theorems for stochastic processes. *Theory of Probability & Its Applications*, 1(3):261–290, 1956.
- [35] W. Whitt. Some useful functions for functional limit theorems. *Mathematics of operations research*, 5(1):67–85, 1980.
- [36] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, 2002.
- [37] W. Whitt. Proofs of the martingale FCLT. *Probab. Surv.*, 4:268–302, 2007.
- [38] Y. P. Yang and C. Knessl. Asymptotic analysis of the M/G/1 queue with a time-dependent arrival rate. *Queueing systems*, 26(1-2):23–68, 1997.